

# A Review of Speaker Diarization: Recent Advances with Deep Learning

Tae Jin Park<sup>a,\*</sup>, Naoyuki Kanda<sup>b,\*</sup>, Dimitrios Dimitriadis<sup>b,\*</sup>, Kyu J. Han<sup>c,\*</sup>, Shinji Watanabe<sup>d,\*</sup>, Shrikanth Narayanan<sup>a</sup>

<sup>a</sup>University of Southern California, Los Angeles, USA

<sup>b</sup>Microsoft, Redmond, USA

<sup>c</sup>ASAPP, Mountain View, USA

<sup>d</sup>Johns Hopkins University, Baltimore, USA

## Abstract

Speaker diarization is a task to label audio or video recordings with classes that correspond to speaker identity, or in short, a task to identify “who spoke when”. In the early years, speaker diarization algorithms were developed for speech recognition on multispeaker audio recordings to enable speaker adaptive processing. These algorithms also gained their own value as a standalone application over time to provide speaker-specific metainformation for downstream tasks such as audio retrieval. More recently, with the emergence of deep learning technology, which has driven revolutionary changes in research and practices across speech application domains, rapid advancements have been made for speaker diarization. In this paper, we review not only the historical development of speaker diarization technology but also the recent advancements in neural speaker diarization approaches. Furthermore, we discuss how speaker diarization systems have been integrated with speech recognition applications and how the recent surge of deep learning is leading the way of jointly modeling these two components to be complementary to each other. By considering such exciting technical trends, we believe that this paper is a valuable contribution to the community to provide a survey work by consolidating the recent developments with neural methods and thus facilitating further progress toward a more efficient speaker diarization.

**Keywords:** speaker diarization, automatic speech recognition, deep learning

## 1. Introduction

“Diarize” means making a note or keeping an event in a diary. Speaker diarization, like keeping a record of events in such a diary, addresses the question of “who spoke when” [1, 2, 3] by logging speaker-specific salient events on multiparticipant (or multispeaker) audio data. Throughout the diarization process, the audio data would be divided and clustered into groups of speech segments with the same speaker identity/label. As a result, salient events, such as non-speech/speech transition, speaker turn change detection or speaker clustering, are automatically performed. In general, this process does not require any prior knowledge of the speakers, such as their real identity or the number of participating speakers in the audio data. Thanks to its innate feature of separating audio streams by these speaker-specific events, speaker diarization can be effectively employed for indexing or analyzing various types of audio data, e.g., audio/video broadcasts from media stations, conversations in conferences, personal videos from online social media or hand-held devices, court proceedings, business meetings, earnings reports in a financial sector, just to name a few.

Traditionally speaker diarization systems consist of multiple, independent sub-modules as presented in Fig. 1. To mitigate any artifacts in acoustic environments, various front-end

processing techniques, for example, speech enhancement, dereverberation, speech separation or target speaker extraction, are employed. Voice or speech activity detection (SAD) is then applied to separate speech from non-speech events. Raw speech signals in the selected speech portion transformed to acoustic features or embedding vectors. In the clustering stage, the speech portions transformed are grouped and labeled by speaker classes and in the post-processing stage, the clustering results are further refined. Each of these sub-modules is optimized individually in general.

### 1.1. Historical Development of Speaker Diarization

During the early years of diarization technology (in the 1990s), the research objective was to benefit automatic speech recognition (ASR) on air traffic control dialogues and broadcast news recordings, by separating each speaker’s speech segments and enabling speaker-adaptive training of acoustic models [4, 5, 6, 7, 8, 9, 10]. In this period some fundamental approaches to the measuring the distance between speech segments for speaker change detection and clustering, such as generalized likelihood ratio (GLR) [4] and Bayesian information criterion (BIC) [11], were developed and quickly became the golden standard. All these efforts collectively laid out paths to consolidate activities across research groups worldwide, leading to several research consortia and challenges in the early 2000s, among which there were the Augmented Multiparty Interaction (AMI) Consortium [12] supported by the European

\*Authors contributed equally

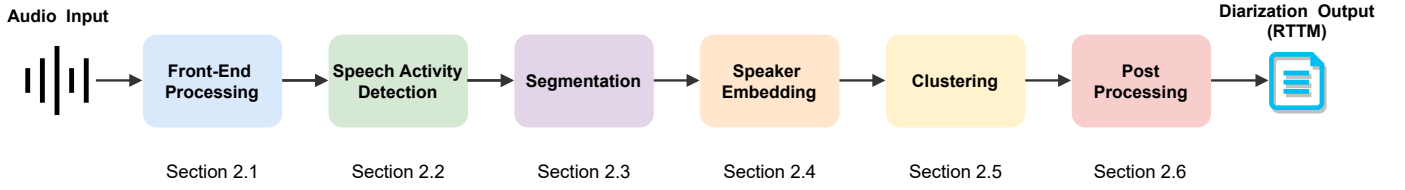


Fig. 1: Traditional speaker diarization system.

Commission and the RT Evaluation [13] hosted by the National Institute of Standards and Technology (NIST). These organizations, spanning over from a few years to a decade, had fostered further advancements on speaker diarization technologies across different data domains from broadcast news [14, 15, 16, 17, 18] and conversational telephone speech (CTS) [19, 20, 21, 22] to meeting conversations [23, 24, 25, 26, 27]. The new approaches resulting from these advancements include, but not limited to, beamforming [28], Information Bottleneck Clustering (IBC) [27], variational Bayesian (VB) approaches [29], joint factor analysis (JFA) [22].

Speaker specific representation in a total variability space derived from simplified JFA, known as i-vector [30], has found great success in speaker recognition and was quickly adopted by speaker diarization systems as feature representation for short speech segments, segmented in an unsupervised fashion. i-vector successfully replaced its predecessors such as merely Mel-frequency cepstral coefficient (MFCC) or speaker factors (or eigenvoices) [31] to bolster clustering performance in speaker diarization, being combined with principal component analysis (PCA) [32, 33], variational Bayesian Gaussian mixture model (VB-GMM) [34], mean shift [35] and probabilistic linear discriminant analysis (PLDA) [36].

Since the advent of deep learning in the 2010s, there has been a considerable amount of research to take advantage of powerful modeling capabilities of the neural networks for speaker diarization. One representative example is the extraction of the speaker embeddings using neural networks, such as the d-vectors [37, 38, 39] or the x-vectors [40], which most often are embedding vector representations based on the bottleneck layer output of a deep neural network (DNN) trained for speaker recognition. The shift from i-vector [30, 33, 41, 42] to these neural embeddings contributed to enhanced performance, easier training with more data [43], and robustness against speaker variability and acoustic conditions. More recently, end-to-end neural diarization (EEND) where individual sub-modules in the traditional speaker diarization systems (c.f., Fig. 1) can be replaced by one neural network gets more attention with promising results [44, 45]. This research direction, although not fully matured yet, could open up unprecedented opportunities to address challenges in the field of speaker diarization, such as, the joint optimization with other speech applications, with overlapping speech, if large-scale data is available for training such powerful network-based models.

## 1.2. Motivation

Till now, there are two well-rounded overview papers in the area of speaker diarization that survey the development of speaker diarization technology with different focuses. In [2], various speaker diarization systems and their subtasks in the context of broadcast news and CTS data are reviewed up to till mid 2000s. Thus, the historical progress of speaker diarization technology development in the 1990s and early 2000s are covered. Contrarily, the focus of [3] was put more on speaker diarization for meeting speech and its respective challenges. This paper thus weighs more in the corresponding technologies to mitigate problems from the perspective of meeting environments, where there are usually more participants than broadcast news or CTS data and multi-modal data is frequently available. Since these two papers were published, speaker diarization systems have gone through a lot of notable changes, especially from the leap-frog advancements in deep learning approaches addressing technical challenges across multiple machine learning domains. We believe that this survey work is a valuable contribution to the community to consolidate the recent developments with neural methods and thus facilitate further progress toward a more efficient diarization.

### 1.3. Overview and Taxonomy of Speaker Diarization

Attempting to categorize the existing, most-diverse speaker diarization technologies, both in the space of modularized speaker diarization systems before the deep learning era and those based on neural networks of recent years, a proper grouping would be helpful. The main categorization we adopt in this paper is based on two criteria, resulting total of four categories, as shown in Table 1. The first criterion is whether the model is trained based on speaker diarization-oriented objective function or not. Any trainable approaches to optimize models in a multispeaker situation and learn relations between speakers are categorized into the “diarization objective” class. The second criterion is whether multiple modules are jointly optimized toward some objective function. If a single sub-module is replaced into a trainable one, such method is categorized into the “Single-module optimization” class. Conversely, joint modeling of segmentation and clustering [43], joint modeling of speech separation and speaker diarization [66] or fully end-to-end neural diarization system [44, 45] is categorized into the “Joint optimization” class.

Note that our intention of this categorization is to help readers to quickly overview the broad development in the field, and it is not our intention to divide the categories into superior-inferior. Also, while we are aware of many techniques that fall into the category “Non-Diarization Objective” and “Joint Opti-

Table 1: Table of Taxonomy

	Trained based on Non-diarization Objective	Trained based on Diarization Objective
<b>Single-module Optimization</b>	<b>Section 2.1–2.6</b> Front-end [46, 47, 48], speaker embedding [49, 50, 40], SAD [51], etc.	<b>Section 3.1</b> Affinity matrix refinement [52], IDEC [53], TS-VAD [54], etc.
<b>Joint Optimization</b>	<b>Section 2.7</b> VB-HMM [55], VBx [56] <b>Out of scope</b> Joint front-end & ASR [57, 58, 59, 60, 61, 62], joint speaker identification & speech separation [63, 64], etc.	<b>Section 3.2</b> UIS-RNN [43], RPN [65], online RSAN [66], EEND [44, 45], etc. <b>Section 4</b> Joint ASR & speaker diarization. [67, 68, 69, 70], etc.

mization” (e.g., joint front-end and ASR [57, 58, 59, 60, 61, 62] and joint speaker identification and speech separation [63, 64]), we exclude them in the paper to focus on the review of speaker diarization techniques.

#### 1.4. Diarization Evaluation Metrics

##### 1.4.1. Diarization Error Rate

The accuracy of speaker diarization system is measured using diarization error rate (DER) [71] where DER is the sum of three different error types: False alarm (FA) of speech, missed detection of speech and confusion between speaker labels.

$$\text{DER} = \frac{\text{FA} + \text{Missed} + \text{Speaker-Confusion}}{\text{Total Duration of Time}}. \quad (1)$$

To establish a one-to-one mapping between the hypothesis outputs and the reference transcript, Hungarian algorithm [72] is employed. In the 2006 RT evaluation [71], 0.25 s of “no score” collar (also referred to as “score collar”) is set around every boundary of reference segment to mitigate the effect of inconsistent notation and human errors in reference transcript and this evaluation scheme has been most widely used in speaker diarization studies.

##### 1.4.2. Jaccard Error Rate

The Jaccard error rate (JER) was first introduced in the DI-HARD II evaluation. The goal of JER is to evaluate each speaker with equal weight. Unlike DER, JER does not use speaker error to obtain the error value.

$$\text{JER} = \frac{1}{N} \sum_i \frac{\text{FA}_i + \text{MISS}_i}{\text{TOTAL}_i}. \quad (2)$$

In Eq. (2), TOTAL is the union of the  $i$ -th speaker’s speaking time in the reference transcript and the  $i$ -th speaker’s speaking time in the hypotheses. The sum of FA and MISS divided by the TOTAL value is then averaged over  $N_{ref}$ -speakers in the reference script. Since JER is using union operation between reference and the hypotheses, JER never exceeds 100%, whereas

DER can sometimes reach way over 100%. DER and JER are highly correlated but if a subset of speakers are dominant in the given audio recording, JER tends to be higher than the ordinary case.

##### 1.4.3. Word-level Diarization Error Rate

While DER is based on the duration of the speaking time of each speaker, word-level DER (WDER) is designed to measure the error that is caused in the lexical (output transcription) side. The motivation of WDER is the discrepancy between DER and the accuracy of the final transcript output since DER relies on the duration of the speaking time that is not always aligned with the word boundaries. The concept of word-breakage ratio was proposed in Silovsky et al. [73] where word-breakage shares similar idea with WDER. Unlike WDER, word-breakage ratio measures the number of speaker-change points occur inside a word boundary. The work in Park and Georgiou [74] suggested the term WDER, evaluating the diarization output with ground-truth transcription. More recently, the joint ASR and speaker diarization system was evaluated in the WDER format in Shafey et al. [67]. Although the way of calculating WDER would differ over the studies, but the underlying mechanism is that the diarization error is calculated by counting the correctly or incorrectly labeled words.

#### 1.5. Paper Organization

The rest of the paper is organized as follows.

- In Section 2, we overview techniques belonging to the “Non-diarization objective” class in the proposed taxonomy, mostly those used in the traditional, modular speaker diarization systems. While there are some overlaps with the counterpart sections of the aforementioned two survey papers [2, 3] in terms of reviewing notable developments in the past, this section would add more latest schemes in the corresponding components of the speaker diarization systems.

- In Section 3, we discuss advancements mostly leveraging DNNs trained with the diarization objective where single sub-modules are independently optimized (Subsection 3.1) or jointly optimized (Subsection 3.2) toward fully end-to-end speaker diarization.
- In Section 4, we present a perspective of how speaker diarization has been investigated in the context of ASR, reviewing historical interactions between these two domains to peek into the past, present and future of speaker diarization applications.
- Section 5 provides information on speaker diarization challenges and corpora to facilitate research activities and anchor technology advances. We also discuss evaluation metrics such as DER, JER and Word-level DER (WDER) in this section.
- We share a few examples of how speaker diarization systems are employed in both research and industry practices in Section 6 and conclude this work in Section 7, providing summary and future challenges in speaker diarization.

## 2. Modular Speaker Diarization Systems

This section provides an overview of algorithms for speaker diarization belonging to the “Non-diarization objective” class, as shown in Table 1. Each subsection in this section corresponds to the explanation of each module in the traditional speaker diarization system, as shown in Figure 1. In addition to the introductory explanation of each module, this section also summarizes the recent techniques within the module.

### 2.1. Front-end Processing

This section describes mostly front-end techniques, used for speech enhancement, dereverberation, speech separation, and speech extraction as part of the speaker diarization pipeline. Let  $s_{i,f,t} \in \mathbb{C}$  be the short-time Fourier Transform (STFT) representation of the source speaker  $i$  on the frequency bin  $f$  at frame  $t$ . The observed noisy signal  $x_{t,f}$  can be represented by a mixture of the source signals, a room impulse response  $h_{i,f,t} \in \mathbb{C}$ , and additive noise  $n_{t,f} \in \mathbb{C}$ ,

$$x_{t,f} = \sum_{i=1}^K \sum_{\tau} h_{i,f,\tau} s_{i,f,t-\tau} + n_{t,f}, \quad (3)$$

where  $K$  denotes the number of speakers present in the audio signal.

The aim of the front-end techniques described in this section is to estimate the original source signal  $\hat{\mathbf{x}}_{i,t}$  given the observation  $\mathbf{X} = (\{x_{t,f}\}_f)_t$  for the downstream diarization task,

$$\hat{\mathbf{x}}_{i,t} = \text{FrontEnd}(\mathbf{X}), \quad i = 1, \dots, K, \quad (4)$$

where  $\hat{\mathbf{x}}_{i,t} \in \mathbb{C}^D$  denotes the  $i$ -th speaker’s estimated STFT spectrum with  $D$  frequency bins at frame  $t$ .

Although there are numerous speech enhancement, dereverberation, and separation algorithms, e.g., [75, 76, 77], herein most of the recent techniques used in the DIHARD challenge series [78, 79, 80], LibriCSS meeting recognition task [81, 82], and the CHiME-6 challenge track 2 [83, 84, 85] are covered.

#### 2.1.1. Speech Enhancement and Denoising

Speech enhancement techniques focus mainly on the suppression of the noise component of the noisy speech, which has shown a significant improvement thanks to deep learning. For example, long short-term memory (LSTM)-based speech enhancement [86, 87] is used as a front-end technique in the DIHARD II baseline [79], i.e.,

$$\hat{\mathbf{x}}_t = \text{LSTM}(\mathbf{X}), \quad (5)$$

where we only consider the single source example (i.e.,  $K = 1$ ) and omit the source index  $i$ . This is a regression-based approach by minimizing the objective function,

$$\mathcal{L}_{\text{MSE}} = \|\mathbf{s}_t - \hat{\mathbf{x}}_t\|^2. \quad (6)$$

The log power spectrum or ideal ratio mask is often used as the target domain of the output  $\mathbf{s}_t$ . Also, the speech enhancement used in [88] applies this objective function in each layer based on a progressive manner.

The effectiveness of the speech enhancement techniques can be enhanced via multichannel processing, including minimum variance distortionless response (MVDR) beamforming [75]. [82] demonstrates the significant improvement of the DER from 18.3% to 13.9% in the LibriCSS meeting task based on mask-based MVDR beamforming [89, 90].

#### 2.1.2. Dereverberation

Compared with other front-end techniques, the major dereverberation techniques used in various tasks are based on statistical signal processing methods. One of the most widely used techniques is the weighted prediction error (WPE) based dereverberation [91, 92, 93].

The basic idea of WPE, for the case of single source, i.e.,  $K = 1$ , without noise, is to decompose the original signal model Eq. (3) into the early reflection  $x_{t,f}^{\text{early}}$  and late reverberation  $x_{t,f}^{\text{late}}$  as follows:

$$x_{t,f} = \sum_{\tau} h_{f,\tau} s_{f,t-\tau} = x_{t,f}^{\text{early}} + x_{t,f}^{\text{late}}. \quad (7)$$

WPE tries to estimate filter coefficients  $\hat{h}_{f,t}^{\text{wpe}} \in \mathbb{C}$ , which maintain the early reflection while suppressing the late reverberation based on the maximum likelihood estimation.

$$\hat{x}_{t,f}^{\text{early}} = x_{t,f} - \sum_{\tau=\Delta}^L \hat{h}_{f,\tau}^{\text{wpe}} x_{f,t-\tau}, \quad (8)$$

where  $\Delta$  denotes the number of frames to split the early reflection and late reverberation, and  $L$  denotes the filter size.

WPE is widely used as one of the gold standard front-end processing methods, e.g., it is part of the DIHARD and CHiME,



which are both the baseline and the top-performing systems [78, 79, 80, 83, 84]. Although the performance improvement of WPE-based dereverberation is not significant, it provides solid performance improvement across almost all tasks. Moreover, WPE is based on linear filtering and since it does not introduce signal distortions, it can be safely combined with downstream front-end and back-end processing steps. Similar to the speech enhancement techniques, WPE-based dereverberation demonstrates additional performance improvements when applied on multichannel signals.

### 2.1.3. Speech Separation

Speech separation is a promising family of techniques when the overlapping speech regions are significant. The effectiveness of multichannel speech separation based on beamforming has been widely confirmed [28, 94, 95]. For example, in the CHiME-6 challenge [83], guided source separation (GSS) [95] based multichannel speech extraction techniques have been used to achieve the top result. On the other hand, single-channel speech separation techniques [46, 47, 48] do not often show any significant effectiveness in realistic multispeaker scenarios, such as the LibriCSS [81] or the CHiME-6 tasks [83], where speech signals are continuous and contain both overlapping and overlap-free speech regions. The single-channel speech separation systems often produce a redundant non-speech or even a duplicated speech signal for the non-overlap regions, and as such the “leakage” of audio causes many false alarms of speech activity. A leakage filtering method was proposed in [96] tackle the problem, where a significant improvement in the diarization performance was observed after including this processing step in the top-ranked system on the VoxCeleb Speaker Recognition Challenge 2020 [97].

### 2.2. Speech Activity Detection

SAD, also known as voice activity detection (VAD), distinguishes speech from non-speech such as background noise. SAD plays a significant role not only in speaker diarization but also in speaker recognition and speech recognition systems since SAD is a pre-processing step that can create errors that propagate through the whole speech recognition or speaker recognition pipeline. A SAD system consists mostly of two major parts. The first one is a feature extraction front-end, where acoustic features such as MFCCs are extracted. The other part is a classifier, where a model predicts and decides whether the input frame is a speech or not.

The SAD techniques are usually variants of decision rules based on the different features. These features may include zero crossing rate [98], pitch [99], signal energy [100], and higher order statistics in the linear predictive coding residual domain [101]. In terms of classification model, a SAD system based on statistical models [102] on spectrum, Gaussian mixture models (GMMs) [103] and on Hidden Markov Models (HMMs) [104, 105] has been widely used.

After the deep learning approaches gained popularity in the speech signal processing field, numerous DNN-based SAD systems were proposed. An MLP-based SAD with a context window proposed in [106] showed improved performances

compared with the traditional GMM-based SAD [103], which was tested in challenging environments from webvideos such as singing voice, music and noisy recordings. A convolutional neural network (CNN) based SAD system was proposed in [107] demonstrated about 10%-25% relative improvement compared with the MLP-based SAD system. An in-depth study on recurrent neural network (RNN)-based SAD systems was conducted in [108] and the proposed coordinated-gate (CG-LSTM) model demonstrated superior performance to the multilayer perceptron (MLP)-based SAD system. More recently, an attention model was applied to a SAD system [109], which outperformed the LSTM and MLP models.

The performance of SAD largely affects the overall performance of the speaker diarization system as it can create a significant amount of false positive salient events or miss speech segments [110]. A common practice in speaker diarization tasks is to report DER with “oracle SAD” setup which indicates that the system output is using SAD output that is identical to the ground truth. Conversely, the system output with an actual speech activity detector is referred to as “system SAD” output.

### 2.3. Segmentation

In the context of speaker diarization, speech segmentation is a process of breaking the input audio stream into multiple segments to obtain speaker-uniform segments. Therefore, the output unit of the speaker diarization system is determined via a segmentation process. In general, speech segmentation methods for speaker diarization are divided into two major categories: Segmentation by speaker-change point detection and uniform segmentation. Segmentation by the speaker-change point detection was the gold standard of the earlier speaker diarization systems, where speaker-change points are detected by comparing two hypotheses:  $H_0$  assumes that both the left and right speech windows are from the same speaker, whereas  $H_1$  assumes that the two speech windows are from the different speakers. To test these two hypotheses, metric-based approaches [111, 112] were most widely applied. In metric-based approaches, the similarity score between the two consecutive speech windows is calculated by modeling the speech feature with a stochastic process such as a Gaussian process and applying a metric to the models created from the speech feature. Let  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  be the sequence of speech features extracted from the given audio stream;  $\mathbf{x}$  is drawn from an independent multivariate Gaussian process:

$$\mathbf{x}_i \sim N(\mu_i, \Sigma_i), \quad (9)$$

where  $\mu_i, \Sigma_i$  is the mean vector and covariance matrix of the  $i$ -th feature window. Using the multivariate Gaussian process assumption, the two hypotheses  $H_0$  and  $H_1$  can be set up as follows:

$$\begin{aligned} H_0 : \mathbf{x}_1 \cdots \mathbf{x}_N &\sim N(\mu, \Sigma), \\ H_1 : \mathbf{x}_1 \cdots \mathbf{x}_i &\sim N(\mu_1, \Sigma_1), \\ &\mathbf{x}_{i+1} \cdots \mathbf{x}_N \sim N(\mu_2, \Sigma_2), \end{aligned} \quad (10)$$

where  $H_0$  models two sample windows with one Gaussian which means there is no speaker-change point, whereas  $H_1$

models two sample windows with two Gaussians, which means there is a speaker-change point. A slew of criteria for the metric-based approach were proposed to quantify the likelihood of the two hypotheses. The examples include the Kullback-Leibler (KL) distance [113], Generalized Likelihood Ratio (GLR) [114, 115, 116] and BIC [111, 117]. Among these criteria for threshold approaches, BIC has been the most widely used method followed by numerous variants [118, 119, 120, 121]. Thus, in this section, we introduce BIC as a representative example for a metric-based method. If we apply BIC and to the hypotheses described in Eq. (10), a BIC value between two models from two hypotheses is expressed as follows:

$$BIC(i) = N \log |\Sigma| - N_1 \log |\Sigma_1| - N_2 \log |\Sigma_2| - \lambda P, \quad (11)$$

where the sample covariance  $\Sigma$  is from  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ ,  $\Sigma_1$  is from  $\{\mathbf{x}_1, \dots, \mathbf{x}_i\}$  and  $\Sigma_2$  is from  $\{\mathbf{x}_{i+1}, \dots, \mathbf{x}_N\}$  and  $P$  is the penalty term [111] defined as

$$P = \frac{1}{2} \left( d + \frac{1}{2} d(d+1) \right) \log N, \quad (12)$$

where  $d$  denotes the dimension of the feature;  $N_1$  and  $N_2$  are frame lengths of each window, respectively and  $N = N_1 + N_2$ . The penalty weight  $\lambda$  is generally set to  $\lambda = 1$ . The change point is set when the following equation becomes true:

$$\left\{ \max_i BIC(i) \right\} > 0. \quad (13)$$

In general, if speech segmentation is done using the speaker-change point detection method, the length of each segment is not consistent. Therefore, after the advent of the i-vector [30] and DNN-based embeddings [49, 40], the segmentation based on speaker-change point detection was mostly replaced with uniform segmentation [35, 122, 39], since varying lengths of the segment created an additional variability into the speaker representation and deteriorated the fidelity of the speaker representations. In uniform segmentation schemes, the given audio stream input is segmented with a fixed window length and overlap length. Thus, the unit duration of the speaker diarization output stays constant. However, the process of uniform segmentation of the input signals for diarization poses some potential problems because it introduces a trade-off error related to the segment length. The segments created from the uniform segmentation need to be sufficiently short to safely assume that they do not contain multiple speakers. However, at the same time it is important to capture sufficient acoustic information to extract reliable speaker representations.

#### 2.4. Speaker Representations and Speaker Embeddings

In this section, we explain a few popular methods for measuring the similarity of speech segments. These methods are often paired with certain clustering algorithms, which will be explained in the next section. We first introduce metric-based approaches for similarity measures which were popular from the late 1990s to early 2000s. Metric-based approaches are usually employed together with the segmentation approaches

based on a speaker-change point detection. We then introduce widely used speaker representations for speaker diarization systems that are usually employed together with the uniform segmentation method described in Section 2.4.2 and Section 2.4.3.

##### 2.4.1. Metric Based Similarity Measure

From the late 1990s to early 2000s, metric-based approaches were most commonly used for measuring the similarity of speech segments for the speaker diarization systems. The metric-based methods that are used for segmentation are also used for comparing the distances to find the closest cluster. The bottom-up approach, also known as agglomerative hierarchical clustering (AHC), was the most widely used method with the metric-based similarity to perform speaker diarization [3]. In the bottom-up approach, the clustering process starts with one node, which is equivalent to a cluster, per speech segment. Then, this approach iteratively chooses the closest matching clusters based on the metric-based similarity to merge. The total number of the clusters is gradually reduced until the closest metric-based similarity reaches a stopping criterion. For the metrics for similarity computation, the methods applied for metric-based speaker segmentation are also applied to clustering: KL distance [113], GLR [114, 115, 116], and BIC [111, 117]. As in the metric-based segmentation, the BIC method was also the most extensively used metric due to its effectiveness and ease of implementation. The BIC approach for measuring similarity among clusters can be described as follows. For the given nodes to be clustered,  $\mathcal{S} = \{s_1, \dots, s_k\}$ , BIC method models each node  $s_i$  as a multivariate Gaussian distribution  $N(\mu_i, \Sigma_i)$  where  $\mu_i$  and  $\Sigma_i$  denote the mean vector and covariance matrix of the merged samples in the node  $s_i$ . On the other hand,  $\mu$  and  $\Sigma$  are mean vector and covariance matrix of the merged cluster of  $s_1$  and  $s_2$ . Thus, the similarity metric-based on BIC for merging the nodes  $s_1$  and  $s_2$  is calculated as follows:

$$BIC = n \log |\Sigma| - n_1 \log |\Sigma_1| - n_2 \log |\Sigma_2| - \lambda P, \quad (14)$$

where the  $\lambda$  and  $P$  values,  $N_1$  and  $N_2$  are cluster size of  $s_1$  and  $s_2$ , respectively, and  $n = n_1 + n_2$ . During the clustering process, we keep merging the pairs that have the biggest BIC value until the clustering system meets the stopping criterion. The metric-based approach with bottom-up clustering was popularly used until i-vector and DNN-based speaker representations dominated the speaker diarization research scene.

##### 2.4.2. Joint Factor Analysis, i-vector and PLDA

Before the advent of speaker representations such as i-vector [30] or x-vector [40], Gaussian Mixture Model and Universal Background Model (GMM-UBM) [123] applied to acoustic features demonstrated success in speaker verification tasks. A UBM is a large GMM (typically with 512 to 2048 mixtures) trained to represent the speaker-independent distribution of acoustic features. Thus, a GMM-UBM model can be described with the following quantities: mixture weights, mean values and covariance matrix of the mixtures. Despite the

success on modeling the speaker identity, GMM-UBM based speaker verification systems have suffered from intersession variability [124], which is the variability exhibited by a given speaker from one recording session to another. Such difficulty occurs because the relevance maximum a posteriori (MAP) adaptation step during the speaker enrollment process in the GMM-UBM based speaker verification systems not only captures the speaker-specific characteristics of the speech, but also unwanted channel noise and other nuisances from the acoustic environment.

Joint factor analysis (JFA) [124, 125] was proposed to compensate for the variability issues by separately modeling the inter-speaker variability and the channel or session variability. The JFA approach employs a GMM supervector, which is a concatenated mean of the adapted GMM. For example, suppose a  $F$  by 1 speaker-independent GMM mean vector  $m_c$ , where  $c$  is the mixture component index and  $F$  is the dimension of the feature. Then, a supervector  $\mathbf{M}$  has dimension of  $CF$  by 1 by concatenating the  $F$ -dimensional mean vector for  $C$  mixture components. Thus, the supervector  $\mathbf{M}$  can be described as follows:

$$\mathbf{M} = [m_1^t, m_2^t, \dots, m_C^t]^t. \quad (15)$$

In the JFA approach, the given GMM supervector is decomposed into speaker independent, speaker dependent, channel dependent, and residual components. Thus, the ideal speaker supervector  $\mathbf{M}_J$  can be decomposed as indicated in Eq. (16), where  $\mathbf{m}_J$  denotes a speaker independent supervector from the UBM,  $\mathbf{V}$  denotes a speaker dependent component matrix,  $\mathbf{U}$  denotes a channel dependent component matrix, and  $\mathbf{D}$  denotes a speaker-dependent residual component matrix. Along with these component matrices, vector  $\mathbf{y}$  is for the speaker factors, vector  $\mathbf{x}$  is for the channel factors, and vector  $\mathbf{z}$  is for the speaker-specific residual factors. All of these vectors have a prior distribution of  $N(0, 1)$ .

$$\mathbf{M}_J = \mathbf{m}_J + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}. \quad (16)$$

The JFA approach was followed by the study in [30], in which it was discovered that channel factors in the JFA also contain information regarding the speakers. Thus, Dehak et al. [30] proposed a new method combining the channel and speaker spaces into a combined variability space through a *total variability* matrix. Thus, the total variability matrix  $\mathbf{T}$  models both the channel and the speaker variability, and the latent variable  $\mathbf{w}$  weights the column of the matrix  $\mathbf{T}$ . The variable  $\mathbf{w}$  is referred to as the *i-vector* and is also considered a speaker representation vector. Each speaker and channel in a GMM supervector  $\mathbf{M}_J$  can be modeled as follows:

$$\mathbf{M}_J = \mathbf{m}_J + \mathbf{T}\mathbf{w}, \quad (17)$$

where  $\mathbf{m}_J$  is a speaker-independent and channel-independent supervector that can be taken as a UBM supervector. The process of extracting an i-vector  $\mathbf{w}$  for the given recording is formulated as a MAP estimation problem [126, 30] using the

Baum–Welch statistics extracted using the UBM, mean super-vector  $\mathbf{m}_J$ , and total variability matrix  $\mathbf{T}$  trained from the EM algorithm as parameters. The idea behind a speaker representation was greatly popularized through the use of i-vectors, where the speaker representation vector can contain a numerical feature characterizing the vocal tract of each speaker. The i-vector speaker representations have been employed in not only speaker recognition studies but also in numerous speaker diarization studies [35, 36, 127] and have shown a superior performance compared to metric-based methods such as BIC, GLR, and KL, as mentioned in the previous subsection.

Intersession variability in the i-vector approach has been further compensated using backend procedures, such as a linear discriminant analysis (LDA) [128, 129] and within-class covariance normalization (WCCN) [130, 131], followed by simple cosine similarity scoring. The cosine similarity scoring was later replaced with a probabilistic LDA (PLDA) model in [132]. In the following studies [133, 134], a method applying a Gaussianization of the i-vectors and thus generating Gaussian assumptions in the PLDA, referred to as G-PLDA or simplified PLDA, was proposed for speaker verification. In general, PLDA employs the following modeling for the given speaker representation  $\phi_{ij}$  of the  $i$ -th speaker and  $j$ -th session as indicated below:

$$\phi_{ij} = \mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}. \quad (18)$$

Here,  $\mu$  is the mean vector,  $\mathbf{F}$  is the speaker variability matrix,  $\mathbf{G}$  is the channel variability matrix, and  $\epsilon$  is a residual component. In addition,  $\mathbf{h}_i$  and  $\mathbf{w}_{ij}$  are latent variables for  $\mathbf{F}$  and  $\mathbf{G}$ , respectively. In G-PLDA, both latent variables,  $\mathbf{h}_i$  and  $\mathbf{w}_{ij}$ , are assumed to follow a standard Gaussian prior. During the training process of the PLDA, the mean vector  $\mu$ , a covariance matrix of the residual component  $\Sigma$ , and variability matrices  $\mathbf{F}$  and  $\mathbf{G}$  are estimated using the expectation maximization (EM) algorithm. Based on the estimated variability matrices and the latent variables  $\mathbf{h}_i$  and  $\mathbf{w}_{ij}$ , two hypotheses are tested: hypothesis  $H_0$  for a case in which two samples are from the same speaker, and hypothesis  $H_1$  for when two samples are from different speakers. Under the hypothesis  $H_0$ , the given speaker representations  $\phi_1$  and  $\phi_2$  are modeled as follows with a common latent variable  $\mathbf{h}_i$  for the speaker variability matrix  $\mathbf{F}$ :

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} + \begin{bmatrix} \mathbf{F} & \mathbf{G} & 0 \\ \mathbf{F} & 0 & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{12} \\ \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}. \quad (19)$$

On the other hand, under the hypothesis  $H_1$ ,  $\phi_1$  and  $\phi_2$  are modeled as follows with separate latent variable  $\mathbf{h}_1$  and  $\mathbf{h}_2$  for the speaker variability matrix  $\mathbf{F}$ :

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} + \begin{bmatrix} \mathbf{F} & \mathbf{G} & 0 & 0 \\ 0 & 0 & \mathbf{F} & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{w}_1 \\ \mathbf{h}_2 \\ \mathbf{w}_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}. \quad (20)$$

Using the models introduced above, PLDA calculates the log likelihood by emphasizing the subspace  $\mathbf{F}$ , where the speaker

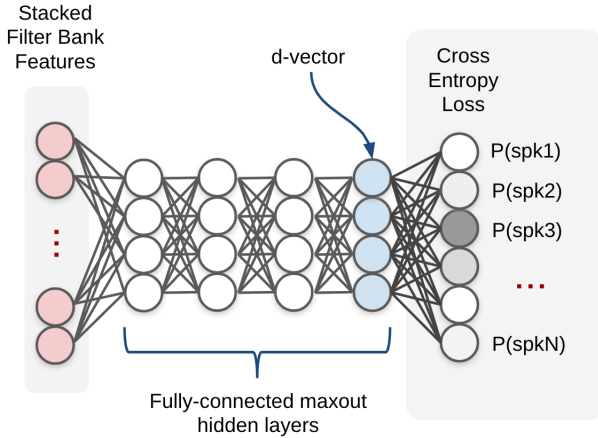


Fig. 2: Diagram of d-vector model.

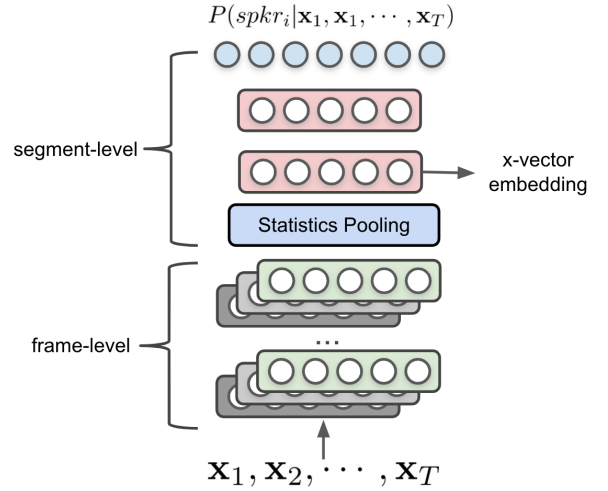


Fig. 3: Diagram of x-vector model.

information covary the most while de-emphasizing the subspace  $\mathbf{G}$  pertaining to the channel variability [135]. In G-PLDA, taking a fixed-length vector  $\phi$  as an input, it is assumed that speaker representation  $\phi$  is generated from a Gaussian distribution. Therefore, the conditional density for a speaker representation vector  $\phi$  is given by the following equation [135]:

$$p(\phi | \mathbf{h}, \mathbf{w}) = \mathcal{N}(\phi | \mu + \mathbf{Fh} + \mathbf{Gw}, \Sigma), \quad (21)$$

where  $\mathcal{N}()$  denotes normal density function. Using the hypotheses in Eq. (19), Eq. (20) and the conditional density in Eq. (21), the log likelihood ratio can be described as follows:

$$s(\phi_1, \phi_2) = \log p(\phi_1, \phi_2 | H_0) - \log p(\phi_1, \phi_2 | H_1). \quad (22)$$

The log likelihood ratio  $s(\phi_1, \phi_2)$  in the above equation was originally used for speaker verification to choose a hypothesis between  $H_0$  and  $H_1$  by checking whether  $s(\phi_1, \phi_2)$  is positive or negative. However, in speaker diarization, the log likelihood  $s(\phi_1, \phi_2)$  is used to check the similarity between clusters. Further details regarding the clustering approach using PLDA is described in Section 2.5.1.

#### 2.4.3. Neural Network Based Speaker Representations

Speaker representations for speaker diarization have also been heavily affected by the rise of deep learning approaches. The idea behind DNN-based representation learning was first introduced for face recognition tasks [136, 137]. As the fundamental idea of a neural network-based representation, we can use the deep neural network architecture to map the input signal source (an image or an audio clip) to a dense vector containing floating-point numbers. This is achieved by taking the values from a layer in the neural network model after forward-propagating the input signal to the layer that we take the values from. The mapping process from the input signal to the speaker embedding is based on the nonlinear modeling capability of multiple layers in the DNNs. In so doing, the

training process of the DNNs allows the neural networks to learn the mapping without specifying any components or factors, which is in contrast to traditional factor analysis models based on decomposable components. In this sense, the components in JFA are more explainable than the parameters trained in DNN models trained for speaker embedding extraction. In addition, DNN-based speaker representation learning does not involve predefined probabilistic models (e.g., GMM-UBM) for the input acoustic features. In relation to this, DNN-based speaker representation achieves an improved efficiency during the inference phase because the solution used by factor-analysis based methods involves a computationally intensive matrix inversion operation [135], whereas DNN-based embedding extractors involve fewer demanding operations, such as multiple linear transformations, with non-linear function computations for obtaining the speaker representation vector. Thus, the representation learning process has become more straightforward and the inference speed has been improved compared to the traditional factor-analysis based methods. Among many of the neural-network based speaker representations, d-vector [49] remains one of the most prominent speaker representation extraction frameworks. The stacked filter bank features, which include context frames as an input feature, are employed, and multiple fully connected layers are trained with cross entropy loss. Speaker representation vectors, also referred to as d-vectors, are obtained from the last fully connected layer, as indicated in Fig. 2. The d-vector scheme appears in numerous speaker diarization papers, e.g., in [39, 43].

DNN-based speaker representations are even more improved when using an x-vector [50, 40], which demonstrates a superior performance, winning the NIST speaker recognition challenge [138] and the first DIHARD challenge [78]. Fig. 3 shows the structure of an x-vector extractor. The time-delay architecture and statistics pooling layer differentiate the x-vector architecture from that of a d-vector. The statistics pooling layer aggregates the frame-level outputs from the previous layer and



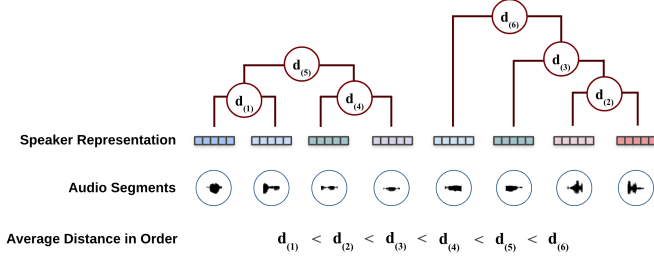


Fig. 4: Agglomerative Hierarchical Clustering.

computes its mean and standard deviation, passing them on to the following layer. Thus, it can allow the extraction of x-vectors from a variable length input. This is advantageous not only for speaker verification but also for speaker diarization because speaker diarization systems are bound to process segments that are shorter than the predetermined uniform segment length when the segment should be truncated at the end of an utterance.

## 2.5. Clustering

As can be seen in Fig. 1, a clustering algorithm is applied to make clusters of the speech segments after generating the speaker embedding for each segment. We introduce the most commonly used clustering methods for speaker diarization.

### 2.5.1. Agglomerative Hierarchical Clustering

AHC is a clustering method that has been constantly employed in many speaker diarization systems with different distance metrics such as BIC [111, 139], KL [140] and PLDA [78, 84, 141]. AHC is an iterative process of merging the existing clusters until the clustering process meets a criterion. The AHC process starts with the calculation of the similarity between  $N$  singleton clusters. At each step, a pair of clusters that has the highest similarity is merged. The iterative merging process of AHC is illustrated in a dendrogram, which is presented in Fig. 4. One of the most important aspects of AHC is the stopping criterion. For the speaker diarization task, the AHC process can be stopped using either a similarity threshold or a target number of clusters. Ideally, if PLDA is used as a distance metric, the AHC process should be stopped at  $s(\phi_1, \phi_2) = 0$  in Eq.(22). However, the stopping metric is adjusted to obtain an accurate number of clusters based on a development set. Conversely, if the number of speakers is known or estimated using other methods, the AHC process can be stopped when the clusters created by the AHC process reaches the pre-determined number of speakers  $k$ .

### 2.5.2. Spectral Clustering

Spectral clustering is a widely used clustering approach for speaker diarization. While there are many variations, spectral clustering involves the following steps.

- i. Affinity matrix calculation: There are many ways to generate affinity matrix  $\mathbf{A}$  depending on the way the affinity

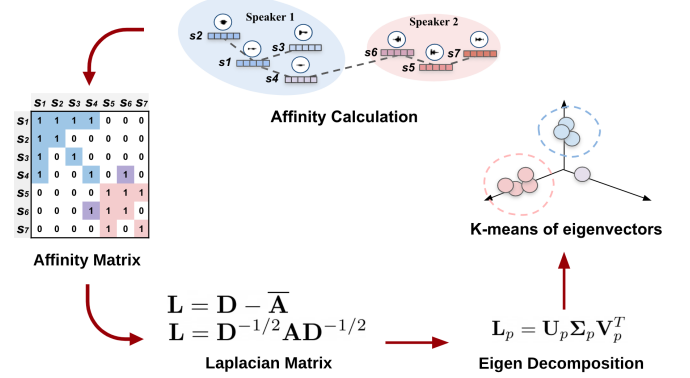


Fig. 5: General steps of spectral clustering.

value is processed. The raw affinity value  $d$  is processed by kernels such as  $\exp(-d^2/\sigma^2)$ , where  $\sigma$  is a scaling parameter. On the other hand, the raw affinity value  $d$  could also be masked by zeroing the values below a threshold to only keep the prominent values.

- ii. Laplacian matrix calculation [142]: The graph Laplacian can be calculated in two ways: normalized and unnormalized. The degree matrix  $\mathbf{D}$  contains diagonal elements  $d_i = \sum_{j=1}^n a_{ij}$  where  $a_{ij}$  is the element of the  $i$ -th row and  $j$ -th column in an affinity matrix  $\mathbf{A}$ .

- (a) Normalized Graph Laplacian:

$$\mathbf{L} = \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}. \quad (23)$$

- (b) Unnormalized Graph Laplacian:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \quad (24)$$

- iii. Eigen decomposition: The graph Laplacian matrix  $\mathbf{L}$  is decomposed into the eigenvector matrix  $\mathbf{X}$  and the diagonal matrix that contains eigenvalues. Thus,  $\mathbf{L} = \mathbf{X} \mathbf{\Lambda} \mathbf{X}^T$ .
- iv. Re-normalization (optional): the rows of  $\mathbf{X}$  is normalized so that  $y_{ij} = x_{ij} / (\sum_j x_{ij}^2)^{1/2}$  where  $x_{ij}$  and  $y_{ij}$  are the elements of the  $i$ -th row and  $j$ -th column in matrices  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively.
- v. Speaker counting: Speaker number is estimated by finding the maximum eigengap [142, 143].
- vi. Spectral embedding clustering: The  $k$ -smallest eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_k$  and the corresponding  $k$  eigenvectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$  are stacked to construct a matrix  $\mathbf{U} \in \mathbb{R}^{n \times k}$ . The row vectors of  $\mathbf{U}$  are referred to as  $k$ -dimensional spectral embeddings. Finally, the spectral embeddings are clustered using a clustering algorithm. In general k-means clustering [144] is employed for clustering the spectral embeddings.

Among many variations of spectral clustering algorithm, the Ng-Jordan-Weiss (NJW) algorithm [145] is often employed for

the speaker diarization task with variation in the kernel for the calculation of the affinity values [146, 147, 33]. Unlike the AHC approach, spectral clustering is mostly used with cosine distance [146, 147, 33, 39, 143]. In addition, the LSTM based similarity measurement with spectral clustering [148] also exhibited competitive performance. Depending on the datasets, the spectral clustering approach with cosine distance measurement outperforms AHC with PLDA [143, 85] while using the same speaker representation for both clustering methods.

### 2.5.3. Other Clustering Algorithms

The k-means algorithm is often employed in studies on speaker diarization [149, 150, 39, 43, 65] due to its simplicity and ease of implementation. However, the k-means algorithm generally underperforms [39, 43] the well-known clustering algorithms such as spectral clustering and AHC. In addition, there are a few speaker diarization studies employed the mean-shift [151] clustering algorithm, which assigns the given data points to the clusters iteratively by finding the modes in a non-parametric distribution. The Mean-shift clustering algorithm was employed in the speaker diarization task with KL distance in [152], i-vector and cosine distance in [35, 153], and i-vector and PLDA in [154].

## 2.6. Post-processing

### 2.6.1. Resegmentation

Resegmentation is a process of refining the speaker boundary that is roughly estimated using the clustering procedure. In [155], the Viterbi resegmentation method based on the Baum-Welch algorithm was introduced. In this method, the estimation of Gaussian mixture model corresponding to each speaker and Viterbi-algorithm-based resgmentation using the estimated speaker GMM are alternately applied.

A method for representing the diarization process based on the variational Bayesian hidden Markov model (VB-HMM) was proposed, and was shown to be superior to Viterbi resegmentation [55, 156]. The VB-HMM-based diarization can be seen as a joint optimization of segmentation and clustering, which will be separately introduced in Section 2.7.

### 2.6.2. System Fusion

As another direction of post processing, there have been a series of studies on the fusion method of multiple diarization results to improve the diarization accuracy. While it is widely known that the system combination generally yields better result for various systems (e.g., speech recognition [157] or speaker recognition [158]), the combination of multiple diarization hypotheses poses several unique problems. First, speaker labeling is not standardized among different diarization systems. Second, the estimated number of speakers may differ among different diarization systems. Finally, the estimated time boundaries may also be different among multiple diarization systems. System combination methods for speaker diarization systems need to handle these problems during the fusion process of multiple hypotheses.

In [159], a method for selecting the best diarization result among many diarization systems was proposed. In this method,

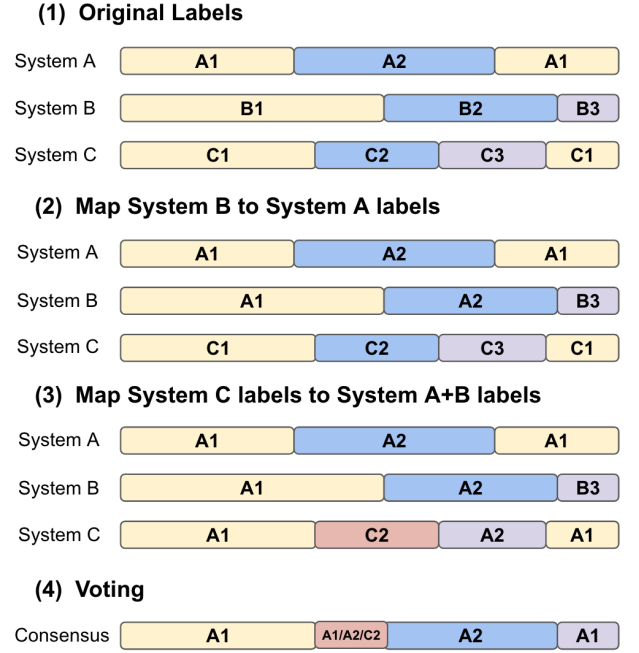


Fig. 6: Example of DOVER system.

a diarization result from each diarization system is treated as one object to be clustered. AHC is applied to the set of diarization results, in which the distance of two clusters is measured using the symmetric DER between the diarization results belonging to the two clusters. The iterative merging process of AHC is executed until the number of clusters becomes two. Finally, in the bigger cluster among the two final clusters, the diarization result that has the smallest distance to all other diarization results is selected as the final result. In [160], two diarization systems are combined by finding the matching between two speaker clusters, and then performing ressegmentation based on the matching result.

More recently, the diarization output voting error reduction (DOVER) method [161] was proposed to combine multiple diarization results based on the voting scheme. In the DOVER method, the speaker labels among different diarization systems are aligned one by one to minimize the DER between the hypotheses (processes 2 and 3 of Fig. 6). After aligning all hypotheses, each system votes its speaker label to each segmented region (each system may have different weights for voting), and the speaker label that gains the highest voting weight is selected for each segmented region (the process 4 of Fig. 6). In case multiple speaker labels get the same voting weight, a heuristic approach is employed to break the ties (such as selecting the result from the first system) is used.

The DOVER method has an implicit assumption that there is no overlapping speech, i.e., at most only one speaker is assigned for each time index. To combine the diarization hypotheses with overlapping speakers, two methods were recently proposed. In [96], the authors proposed the modified DOVER method, in which the speaker labels in different diarization re-

sults are first aligned with a root hypothesis, and the speech activity of each speaker is estimated based on the weighted voting score for each speaker for each small segment. the speakers of multiple hypotheses are aligned via a weighted k-partite graph matching, and the number of speakers  $K$  for each small segment is estimated based on the weighted average of multiple systems to select the top- $K$  voted speaker labels. Both the modified DOVER and DOVER-Lap showed DER improvement for the speaker diarization result with speaker overlaps.

### 2.7. Joint Optimization of Segmentation and Clustering

This subsection introduces a VB-HMM-based diarization technique, which can be regarded as a joint optimization of segmentation and clustering, and thus cannot be well categorized in Section 2.1–2.6. In the VB-HMM framework [55], the speech feature  $\mathbf{X} = (\mathbf{x}_t | t = 1, \dots, T)$  is assumed to be generated from HMM where each HMM state corresponds to one of  $K$  possible speakers. Suppose that we have  $M$  HMM states,  $M$ -dimensional variable  $\mathbf{Z} = (\mathbf{z}_t | t = 1, \dots, T)$  is introduced where  $k$ -th element of  $\mathbf{z}_t$  is 1 if  $k$ -th speaker is speaking at the time index  $t$ , and 0 otherwise. At the same time, the distribution of  $\mathbf{x}_t$  is modeled based on a hidden variable  $\mathbf{Y} = \{\mathbf{y}_k | k = 1, \dots, K\}$ , where  $\mathbf{y}_k$  denotes a low dimensional vector for the  $k$ -th speaker. Given these notations, the joint probability of  $\mathbf{X}$ ,  $\mathbf{Y}$ , and  $\mathbf{Z}$  is decomposed as follows:

$$P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) = P(\mathbf{X} | \mathbf{Z}, \mathbf{Y}) P(\mathbf{Z}) P(\mathbf{Y}), \quad (25)$$

where  $P(\mathbf{X} | \mathbf{Z}, \mathbf{Y})$  is the emission probability modeled by GMM whose mean vector is represented by  $\mathbf{Y}$ ,  $P(\mathbf{Z})$  is the transition probability of the HMM, and  $P(\mathbf{Y})$  is the prior distribution of  $\mathbf{Y}$ . Because  $\mathbf{Z}$  represents the trajectory of speakers, the diarization problem can be expressed as the inference problem of  $\mathbf{Z}$  that maximizes the posterior distribution  $P(\mathbf{Z} | \mathbf{X}) = \int P(\mathbf{Z}, \mathbf{Y} | \mathbf{X}) d\mathbf{Y}$ . Since it is intractable to directly solve this problem, the VB method is used to estimate the model parameters that approximate  $P(\mathbf{Z}, \mathbf{Y} | \mathbf{X})$ .

Recently, a simplified version of VB-HMM that works on the x-vector, known as VBx, was proposed [56, 162]. In VBx,  $P(\mathbf{X} | \mathbf{Z}, \mathbf{Y})$  is calculated using the x-vector based on the PLDA model. While the original VB-HMM works on the granularity of the frame-level feature, VBx works on the granularity of the x-vector, and thus can be seen as a clustering method that jointly models speaker turn and speaker duration.

The VB-HMM diarization was originally designed as a standalone diarization framework. However, it requires parameter initialization to start the VB estimation, and the parameters are usually initialized based on the result of another speaker clustering. In that context, VB-HMM is widely employed as the final step of speaker diarization (e.g., [163, 122, 164]). For example in [164], AHC was first performed to under-cluster the x-vector, and VBx was then applied to obtain a better cluster given the AHC-based result as the initial parameter. Finally, VB-HMM was further applied to refine the boundary obtained by VBx.

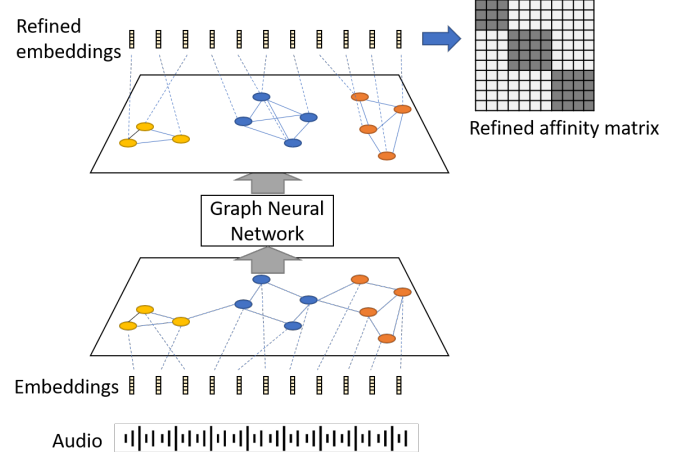


Fig. 7: Speaker diarization with graph neural network

## 3. Recent Advances in Speaker Diarization Using Deep Learning

This section introduces various recent efforts toward deep learning-based speaker diarization techniques. The methods that incorporate deep learning into a single component of speaker diarization, such as clustering or post-processing, are introduced in Section 3.1. The methods that unify several components of speaker diarization into a single neural network are introduced in Section 3.2. For the overview of speaker diarization techniques using deep learning, refer to Table 2. It should be noted that there are some works that take additional input of speaker profiles. These methods may not be categorized as a diarization technique in a traditional definition. Nevertheless, we introduce them as they are optimized in a multispeaker situation to learn the relations between speakers and hence categorized as “Trained Based on the Diarization Objective” in Table 1.

### 3.1. Single-module Optimization

#### 3.1.1. Speaker clustering Enhanced by Deep Learning

Several methods that enhance speaker clustering based on deep learning were proposed. In [52], an approach that purifies the similarity matrix for the spectral clustering based on the graph neural network (GNN) was proposed (Fig. 7). Suppose that a sequence of speaker embeddings  $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$  where  $N$  is the length of the sequence. The first layer of the GNN takes the input  $\{\mathbf{x}_i^0 = \mathbf{e}_i | i = 1, \dots, N\}$ . The GNN then computes the output of the  $p$ -th layer  $\{\mathbf{x}_i^{(p)} | i = 1, \dots, N\}$  as the following:

$$\mathbf{x}_i^{(p)} = \sigma(\mathbf{W} \sum_j \mathbf{L}_{i,j} \mathbf{x}_j^{(p-1)}), \quad (26)$$

where  $\mathbf{L}$  is a normalized affinity matrix added by self-connection,  $\mathbf{W}$  is a trainable weight matrix for the  $p$ -th layer, and  $\sigma(\cdot)$  is a nonlinear function. GNN was optimized by minimizing the distance between the reference affinity matrix and estimated affinity matrix, where the distance was calculated using a combination of histogram loss [177] and nuclear

Table 2: Overview of speaker diarization techniques using deep learning. Note that there are also a lot of studies that use deep learning for front-end, SAD, segmentation, and speaker embedding extraction, which are introduced in Section 2.

Section	Front-end	SAD	Seg.	Speaker embed.	Clust.	Post-proc.	Joint w/ ASR	Speaker profile	Overlap handling	Inference mode	Max # of speakers	References
3.1.1	-	-	-	-	✓	-	-	-	no	offline	Variable	Affinity matrix [52, 165, 166], IDEC [53]
3.1.2	-	-	-	-	✓	-	-	required	no	online	Fixed	RRNN [167]
3.1.3	-*	-	-	-	-	✓	-	required	yes	offline	Fixed	TS-VAD <sup>†</sup> [85, 54]
3.1.3	-*	-	-	-	-	✓	-	-	yes	offline	Variable	EEND-based post-processing <sup>‡</sup> [168]
3.2.1	-	-	✓	-	✓	-	-	-	no	both	Variable	UIS-RNN [43]
3.2.2	-*	-	✓	✓	✓	✓	-	-	yes	offline	Variable	RPN [65]
3.2.3	✓	-	✓	✓	✓	-	-	-	yes	online	Variable	Online-RSAN [66, 169]
3.2.4	-*	✓	✓	✓	✓	-	-	-	yes	offline	Fixed	EEND [44, 45]
3.2.4	-*	✓	✓	✓	✓	-	-	-	yes	offline	Variable	EEND-EDA [170], Speaker-chain [171]
3.2.4	-*	✓	✓	✓	✓	-	-	-	yes	online	Fixed	EEND with speaker tracing buffer [172]
3.2.4	-*	✓	✓	✓	✓	-	-	-	yes	online	Variable	BW-EDA-EEND [173]
3.2.4	-*	✓	✓	✓	✓	-	-	-	yes	offline	Variable	EEND-vector [174]
4.3	-	-	✓	-	✓	-	✓	-	no	offline	Fixed	Speaker-tag insertion [67, 68]
4.3	-*	-	✓	-	✓	-	✓	-	yes	offline	Variable	MAP-decoding [69]
4.3	-*	-	✓	✓	✓	-	✓	required	yes	offline	Variable	End-to-end SA-ASR [175, 176]
4.3	-*	-	✓	✓	-	-	✓	-	yes	offline	Variable	End-to-end SA-ASR+Clustering [70]

<sup>†</sup> TS-VAD can also be interpreted as a joint model of SAD, segmentation, and speaker identification.

<sup>‡</sup> EEND [44, 45] itself is a joint model of SAD, segmentation, speaker embedding, and clustering.

\* It can also be seen that a speech separation module is implicitly embedded in the model to cope with speaker overlaps.

norm. The GNN-based speaker diarization method was evaluated by using the CALLHOME dataset and their in-house meeting dataset and showed significantly better results than the conventional clustering methods. There are also several different approaches to generate the affinity matrix. In [165], a self-attention-based neural network model was introduced to directly generate a similarity matrix from a sequence of speaker embeddings. In [166], several affinity matrices with different temporal resolutions were fused into a single affinity matrix based on a neural network.

As another line of research that improves clustering, a deep-learning-based clustering algorithm, called improved deep embedded clustering (IDEC) was proposed in [53]. The goal is to transform the input features (herein referred to as speaker embeddings) for them to become more separable, given the number of clusters/speakers. The key idea is that each embedding has a probability of “belonging” to each of the available speaker cluster [178, 179], i.e.  $q_{i,j}$  can be interpreted as the probability of assigning sample  $i$  to cluster  $j$  (i.e., a soft assignment):

$$q_{ij} = \frac{(1 + \|z_i - \mu_j\|^2/a)^{-\frac{a+1}{a}}}{\sum_l (1 + \|z_i - \mu_l\|^2/a)^{-\frac{a+1}{a}}}, \quad p_{ij} = \frac{q_{ij}^2/f_i}{\sum_l q_{il}^2/f_l}, \quad (27)$$

where  $z_i$  is the bottleneck feature,  $a$  is the degree of freedom of the Student’s  $t$ -distribution,  $\mu_i$  is the centroid of  $i$ -th cluster and  $f_i$  is the soft cluster frequency with  $f_i = \sum q_{ij}$ . The clusters are iteratively refined based on the target distribution [178] according to the bottleneck features estimated using an autoencoder.

The initial deep embedded clustering (DEC) approach posed some problems. Thus, improved versions of the algorithm have been proposed, where the possibility of trivial (empty) clus-

ters is addressed (under the assumption that the distribution of speaker turns is uniform across all speakers, i.e., all speakers contribute equally to the session). This assumption is not realistic in real meeting environments but it constrains the solution space enough to avoid the empty clusters without affecting the overall performance. An additional loss term penalizes the distance from the centroid  $\mu_i$ , bringing the behavior of the algorithm closer to k-means [53].

Based on these improvements, the loss function of the revisited DEC algorithm consists of three different loss components, i.e.,  $L_c$ , the clustering error term;  $L_r$ , the reconstruction error term of the output of the autoencoder with the input features [179];  $L_u$  is the uniform “speaker airtime” distribution constraint; and  $L_{MSE}$ , the distance of the bottleneck features from the centroids [53],

$$L = \alpha L_c + \beta L_r + \gamma L_u + \delta L_{MSE}, \quad (28)$$

which allows the different loss functions to be weighted differently and the weights  $\alpha, \beta, \gamma$ , and  $\delta$  to be fine-tuned on some held-out data.

### 3.1.2. Learning the Distance Estimator

Data-driven techniques demonstrated good performance in a wide variety of tasks [180]. However, traditional DL architectures can fail when the problem involves relational information between observations [181]. Recently, Relational Recurrent neural networks (RRNN) were introduced by [181, 182, 183] to solve the “relational information learning” task. Speaker diarization is considered to be involved in this class of tasks, since the final decision depends on the distance between speech segments and speaker profiles or centroids.



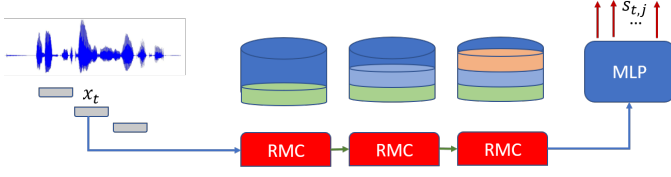


Fig. 8: Continuous speaker identification system based on RMC. The speech signal is segmented uniformly and each segment  $x_t$  is compared against all the available speaker profiles according to a distance metric  $d(\cdot, \cdot)$ . A speaker label  $s_{t,j}$  is assigned to each  $x_t$  minimizing this metric.

The challenges of audio segmentation are detailed in Section 2.3. Thus, speaker embeddings are usually extracted from a network trained to distinguish speakers among thousands of candidates [40]. However, a different level of granularity in the speaker space is required, since the number of participants involved in an interactive meeting scenario is typically small. Aside from this, the distance metric often used is based on a heuristic approach and/or dependent on certain assumptions that do not necessarily hold, e.g., assuming Gaussianity in the case of PLDA [133]. Finally, the audio chunks are treated independently and any temporal context is simply ignored. Most of these issues are addressed using the RRNNs as in [167], where a data-driven, memory-based approach bridges the performance gap between the heuristic and the trainable distance estimating approaches. The use of RRNNs has been successful in several problems requiring relational reasoning [182, 181, 167], and specifically when using the relational memory core (RMC) [181]. Although the task in [167] is speaker ID, the extension to the speaker diarization task is straightforward when the speaker profiles can be pre-estimated, either as centroids using any traditional clustering algorithm, Section 2.5 and 3.1.1, or with prior knowledge.

A novel approach for learning the relationship between the speaker cluster centroids (or speaker profiles) and the embeddings is proposed in [167] (Fig. 8). In this work, the diarization process is considered to be a classification task on an already segmented audio, as in Section 2.3, either uniformly [149] or based on estimated speaker-change salient points [184]. The speaker embeddings  $x_j$  for each segment, assumed to be speaker-homogeneous, are extracted and then compared with all the available speaker profiles or speaker centroids. The most suitable speaker label is assigned to each segment by minimizing the distance-based loss function, i.e., the relationship between embeddings and profiles. As discussed in [167], the RRNN-based distance estimation exhibits consistent improvements in its performance when compared with the more traditional distance estimation approaches such as the cosine distance [30] or the PLDA-based [133] distance.

### 3.1.3. Post Processing Based on Deep Learning

There are a few recent studies on the neural network-based diarization method that is applied on top of the result from a traditional clustering-based speaker diarization. These methods can be categorized as an extension of the post-processing. Medennikov et al. [85, 54] proposed the target-speaker voice

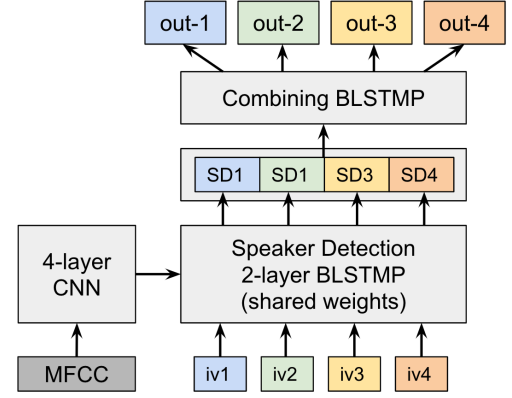


Fig. 9: Target Speaker Voice Activity Detector

activity detection (TS-VAD) to achieve accurate speaker diarization even under noisy conditions with many speaker overlaps. TS-VAD assumes that a set of i-vectors  $\mathcal{E} = \{\mathbf{e}_k \in \mathbb{R}^f | k = 1, \dots, K\}$  are available for each speaker in the audio, where  $f$  is the dimension of i-vector and  $K$  is the number of speakers. As presented in Fig. 9, TS-VAD takes not only a sequence of MFCC,  $\mathbf{X} = (\mathbf{x}_t \in \mathbb{R}^d | t = 1, \dots, T)$ , where  $d$  is the dimension of MFCC and  $T$  the length of the sequence, but also a set of i-vectors  $\mathcal{E}$ . Given  $\mathbf{X}$  and  $\mathcal{E}$ , the model outputs a sequence of  $k$ -dimensional vector  $\mathbf{O} = (\mathbf{o}_t \in \mathbb{R}^K | t = 1, \dots, T)$  where the  $k$ -th element of  $\mathbf{o}_t$  represents the probability of the speech activity of the speaker corresponding to  $\mathbf{e}_k$  at the time frame  $t$ . In other words, the  $k$ -th element of  $\mathbf{o}_t$  is expected to be 1 if the speaker of  $\mathbf{e}_k$  is speaking at time  $t$ , and 0 otherwise.

Because TS-VAD requires the i-vectors of speakers, pre-processing to obtain the i-vectors is necessary. The procedure proposed in [85, 54] is as follows:

1. Apply clustering-based diarization.
2. Estimate i-vectors for each speaker given the diarization result.
3. Repeat (a) and (b).
  - (a) Apply TS-VAD given the estimated i-vectors.
  - (b) Refine i-vectors given the TS-VAD result.

TS-VAD was proposed as a part of the winning system of CHiME-6 Challenge [83], and showed a significantly better DER compared with the conventional clustering based approach [85]. However, it has a drawback, i.e., the maximum number of speakers that the model can handle is limited by the dimension of the output vector.

As a different approach, Horiguchi et al. proposed the application of the EEND model (detailed in Section 3.2.4) to refine the result of a clustering-based speaker diarization [168]. A clustering-based speaker diarization method can handle a large number of speakers but unable to handle overlapped speech. Conversely, EEND has opposite characteristics. To complementarily use the two methods, The authors in [85] first applied a

conventional clustering method. Then, the two-speaker EEND model was iteratively applied for each pair of detected speakers to refine the time boundary of overlapped regions.

### 3.2. Joint Optimization for Speaker Diarization

#### 3.2.1. Joint Segmentation and Clustering

A model called unbounded interleaved-state recurrent neural networks (UIS-RNN) was proposed, which replaced the segmentation and clustering methods with a trainable model [43]. Given the input sequence of embeddings  $\mathbf{X} = (\mathbf{x}_t \in \mathbb{R}^d | t = 1, \dots, T)$ , UIS-RNN generates the diarization result  $\mathbf{Y} = (y_t \in \mathbb{N} | t = 1, \dots, T)$  as a sequence of speaker index for each time frame. The joint probability of  $\mathbf{X}$  and  $\mathbf{Y}$  can be decomposed by the chain rule as follows.

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{x}_1, y_1) \prod_{t=2}^T P(\mathbf{x}_t, y_t | \mathbf{x}_{1:t-1}, y_{1:t-1}). \quad (29)$$

To model the distribution of the speaker change, UIS-RNN then introduces a latent variable  $\mathbf{Z} = (z_t \in \{0, 1\} | t = 2, \dots, T)$ , where  $z_t$  becomes 1 if the speaker indices at time  $t-1$  and  $t$  are different, and 0 otherwise. The joint probability including  $\mathbf{Z}$  is then decomposed as follows.

$$P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = P(\mathbf{x}_1, y_1) \prod_{t=2}^T P(\mathbf{x}_t, y_t, z_t | \mathbf{x}_{1:t-1}, y_{1:t-1}, z_{1:t-1}). \quad (30)$$

Finally, the term  $P(\mathbf{x}_t, y_t, z_t | \mathbf{x}_{1:t-1}, y_{1:t-1}, z_{1:t-1})$  is further decomposed into three components.

$$P(\mathbf{x}_t, y_t, z_t | \mathbf{x}_{1:t-1}, y_{1:t-1}, z_{1:t-1}) = P(\mathbf{x}_t | \mathbf{x}_{1:t-1}, y_{1:t}) P(y_t | z_t, y_{1:t-1}) P(z_t | z_{1:t-1}). \quad (31)$$

Here,  $P(\mathbf{x}_t | \mathbf{x}_{1:t-1}, y_{1:t})$  denotes the sequence generation probability, and modeled by gated recurrent unit (GRU)-based recurrent neural networks.  $P(y_t | z_t, y_{1:t-1})$  denotes the speaker assignment probability, and modeled by a distance-dependent Chinese restaurant process [185], which can model the distribution of unbounded number of speakers. Finally,  $P(z_t | z_{1:t-1})$  represents the speaker change probability and is modeled by the Bernoulli distribution. Since all models are represented by trainable ones, UIS-RNN can be trained in a supervised fashion by finding the parameters that maximize  $\log P(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$  over training data. The inference can be conducted by finding  $\mathbf{Y}$  that maximizes  $\log P(\mathbf{X}, \mathbf{Y})$  given  $\mathbf{X}$  based on the beam search in an online fashion. While UIS-RNN works in an online fashion, it demonstrated better DER than that of the offline system based on spectral clustering.

#### 3.2.2. Joint Segmentation, Embedding Extraction, and Resegmentation

A speaker diarization method based on the region proposal networks (RPN) was proposed to jointly perform segmentation, speaker embedding extraction, and resegmentation [65]. The RPN was originally proposed to detect multiple objects from a two-dimensional image [186], and one-dimensional variant along with the time-axis is used for speaker diarization.

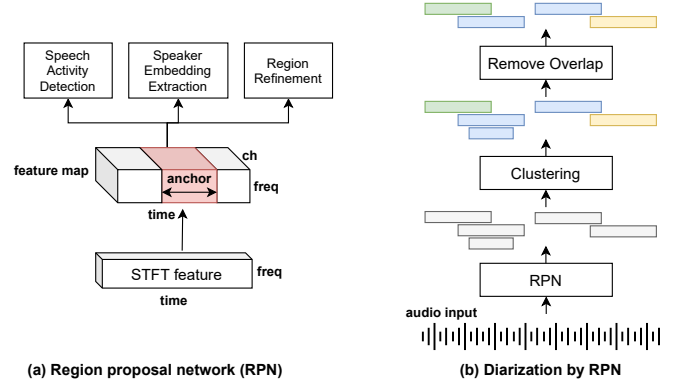


Fig. 10: (a) RPN for speaker diarization, (b) diarization procedure based on RPN.

As can be seen from Fig. 10 (a), the STFT features with a size of time and frequency bin is first converted to the feature map with a size of time, frequency and channels using CNNs. Then, other three types of neural networks are applied on various sizes of sliding windows (named “anchor”) along with the time axis. For each anchor, the three neural networks perform SAD, speaker embedding extraction, and region refinement, respectively. Here, SAD is the task to estimate the probability of speech activity for the anchor region. Speaker embedding extraction is the task to generate an embedding to represent the speaker characteristics of the audio corresponding to the anchor region. Finally, region refinement is the task to estimate the difference between the shape (i.e. duration and center position) of the anchor and that of the corresponding reference region.

The inference procedure by RPN is presented in Fig. 10 (b). RPN is first applied to list the anchors with speech activity probability higher than the pre-determined threshold. The anchors are then clustered using a conventional clustering method (e.g., k-means) based on the speaker embeddings estimated for each anchor. Finally, highly overlapped anchors after region refinement are removed, a method known as the non-maximum suppression.

The RPN-based speaker diarization system has the advantage of handling overlapped speech with possibly any number of speakers. Also, it is much simpler than the conventional speaker diarization system. It was shown in multiple datasets that this system achieved significantly better DER than the conventional clustering-based speaker diarization system [65, 82].

#### 3.2.3. Joint Speech Separation and Diarization

There are also recent researches on the joint modeling of speech separation and speaker diarization. Kounades-Bastian et al. [187, 188] proposed the incorporation of a speech activity model into speech separation based on the spatial covariance model with non-negative matrix factorization. They derived the EM algorithm to estimate separated speech and the speech activity of each speaker from the multichannel overlapped speech. While their method jointly performs speaker diarization and speech separation, their method is based on a statistical mod-

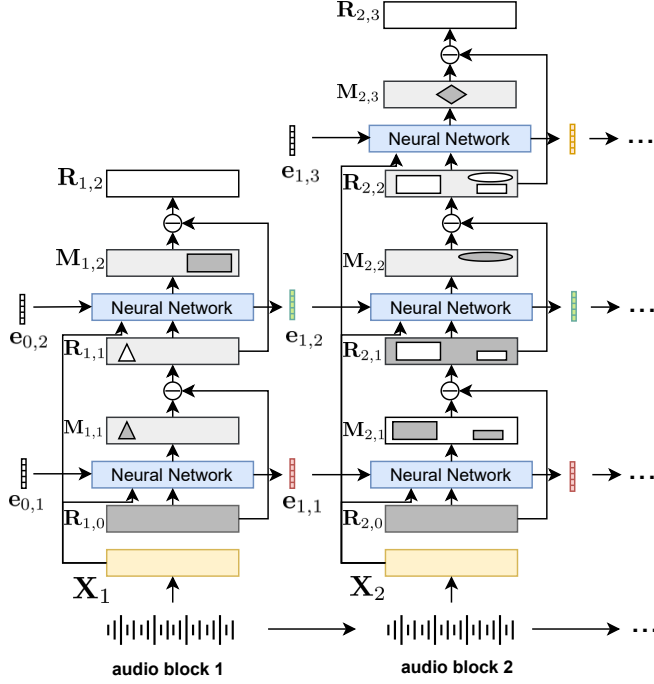


Fig. 11: Joint speech separation, speaker counting, and speaker diarization model.

eling, and estimation was conducted solely based on the observation, i.e. without any model training.

Neumann et al. [66, 169] later proposed a trainable model, namely online Recurrent Selective Attention Network (online RSAN), for joint speech separation, speaker counting, and speaker diarization based on a single neural network (Fig. 11). Their neural network takes the input of spectrogram  $\mathbf{X}_b \in \mathbb{R}^{T \times F}$ , a residual mask  $\mathbf{R}_{b,i-1} \in \mathbb{R}^{T \times F}$ , and a speaker embedding  $\mathbf{e}_{b-1,i} \in \mathbb{R}^d$ , where  $b$  is the index of the audio block;  $i$ , the index of the speaker;  $T$ , the length of the audio block; and  $F$ , the maximum frequency bin of the spectrogram. It outputs the speech mask  $\mathbf{M}_{b,i} \in \mathbb{R}^{T \times F}$  and an updated speaker embedding for the speaker corresponding to  $\mathbf{e}_{b,i}$ . The neural network is applied in an iterative fashion for each audio block  $b$ , and for each speaker  $i$  as follows:

1. Repeat (a) and (b) for  $b = 1, 2, \dots$ 
  - (a)  $\mathbf{R}_{b,0} = \mathbf{1}$
  - (b) Repeat (i)-(iii) for  $i = 1, 2, \dots$  until being stopped at (iii).
    - i.  $\mathbf{M}_{b,i}, \mathbf{e}_{b,i} = \text{NN}(\mathbf{X}_b, \mathbf{R}_{b,i-1}, \mathbf{e}_{b-1,i})$   
( $\mathbf{e}_{b-1,i}$  is set to  $\mathbf{0}$  if it was not calculated previously)
    - ii.  $\mathbf{R}_{b,i} = \max(\mathbf{R}_{b,i-1} - \mathbf{M}_{b,i}, \mathbf{0})$
    - iii. If  $\frac{1}{TF} \sum_{t,f} \mathbf{R}_{b,i}(t, f) < \text{threshold}$ , stop iteration.

A separated speech for speaker  $i$  at audio block  $b$  can be obtained by  $\mathbf{M}_{b,i} \odot \mathbf{X}_b$  where  $\odot$  is the element-wise multiplication. The speaker embedding  $\mathbf{e}_{b,i}$  is used to keep track of the speaker of adjacent blocks. Thanks to the iterative approach,

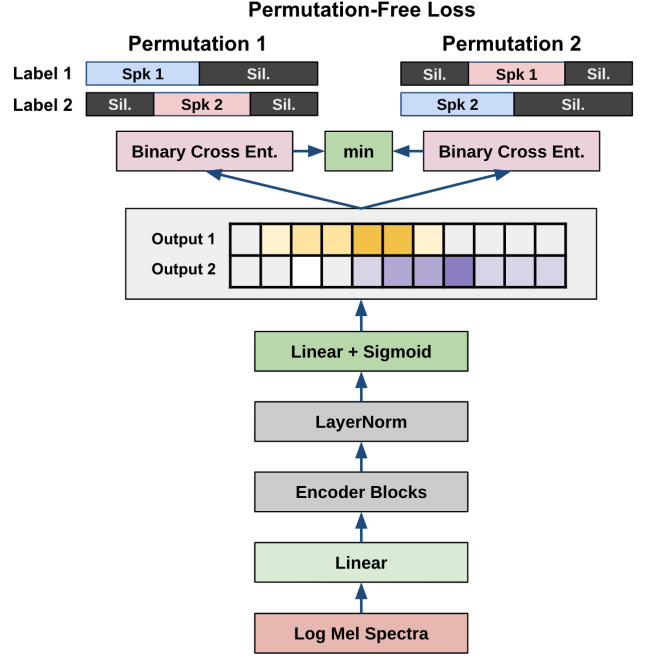


Fig. 12: Two-speaker end-to-end neural diarization model

this neural network can cope with the variable number of speakers while jointly performing speech separation and speaker diarization. The online RSAN was evaluated by using real meeting dataset with up to six speakers, and showed better results than the clustering-based method [169].

### 3.2.4. Fully End-to-end Neural Diarization

Recently, the framework called EEND was proposed [44, 45], which performs all the speaker diarization procedures based on a single neural network. The architecture of EEND is shown in Fig. 12. An input to the EEND model is a  $T$ -length sequence of acoustic features (e.g., log Mel-filterbank),  $\mathbf{X} = (\mathbf{x}_t \in \mathbb{R}^F | t = 1, \dots, T)$ . A neural network then outputs the corresponding speaker label sequence  $\mathbf{Y} = (\mathbf{y}_t | t = 1, \dots, T)$  where  $\mathbf{y}_t = [y_{t,k} \in \{0, 1\} | k = 1, \dots, K]$ . Here,  $y_{t,k} = 1$  represents the speech activity of the speaker  $k$  at the time frame  $t$ , and  $K$  is the maximum number of speakers that the neural network can output. Importantly,  $y_{t,k}$  and  $y_{t,k'}$  can be both 1 for different speakers  $k$  and  $k'$ , indicating that these two speakers  $k$  and  $k'$  are speaking simultaneously (i.e. overlapping speech). The neural network is trained to maximize  $\log P(\mathbf{Y}|\mathbf{X}) \sim \sum_t \sum_k \log P(y_{t,k}|\mathbf{X})$  over the training data by assuming the conditional independence of the output  $y_{t,k}$ . Because there can be multiple candidates of the reference label  $\mathbf{Y}$  by swapping the speaker index  $k$ , the loss function is calculated for all possible reference labels and the reference label that has the minimum loss is used for the error back-propagation, which is inspired by the permutation free objective used in speech separation [47]. EEND was initially proposed using a bidirectional long short-term memory (BLSTM) network [44], and was soon extended to the self-attention-based network [45] by showing the state-of-the-art DER for two-speaker data such as the two-

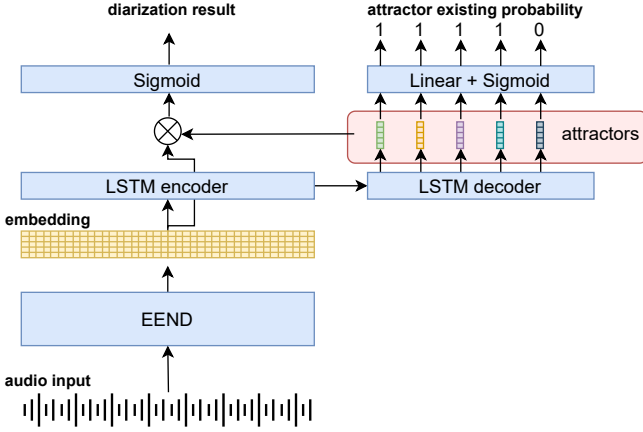


Fig. 13: End-to-end neural diarization with encoder-decoder-based attractor.

speaker excerpt from the CALLHOME dataset (LDC2001S97) and the dialogue audio in the corpus of Spontaneous Japanese [189].

EEND has multiple advantages. First, EEND can handle overlapping speech in a sound way. Second, the network is directly optimized toward the maximization of diarization accuracy, by which we can expect a high accuracy. Third, it can be retrained by a real data (i.e. not synthetic data) just by feeding a reference diarization label while it is often not straightforward for the prior works. However, EEND also has several limitations. First, the model architecture limits the maximum number of speakers that the model can cope with. Second, EEND consists of BLSTM or self-attention based neural networks, making it difficult to do online processing. Third, it was empirically suggested that EEND tends to overfit to the distribution of the training data [44].

To cope with an unbounded number of speakers, several extensions of EEND have been investigated. Horiguchi et al. [170] proposed an extension of EEND with the encoder-decoder-based attractor (EDA) (Fig. 13). This method applies an LSTM-based encoder-decoder on the output of EEND to generate multiple attractors. Attractors are generated until the attractor existing probability becomes less than the threshold. Then, each attractor is multiplied by the embeddings generated from EEND to calculate the speech activity for each speaker. EEND-EDA was evaluated on CALLHOME (two to six speakers) and DIHARD 2 (one to nine speakers) dataset and showed better performance than the clustering-based baseline system. On the other hand, Fujita et al. [171] proposed another approach to output the speech activity one after another by using a conditional speaker chain rule. In this method, a neural network is trained to produce a posterior probability  $P(\mathbf{y}_k | \mathbf{y}_1, \dots, \mathbf{y}_{k-1}, \mathbf{X})$ , where  $\mathbf{y}_k = (y_{t,k} \in \{0, 1\} | t = 1, \dots, T)$  is the speech activity for the  $k$ -th speaker. Then, the joint speech activity probability of all speakers can be estimated from the following speaker-wise conditional chain rule as:

$$P(\mathbf{y}_1, \dots, \mathbf{y}_K | \mathbf{X}) = \prod_{k=1}^K P(\mathbf{y}_k | \mathbf{y}_1, \dots, \mathbf{y}_{k-1}, \mathbf{X}). \quad (32)$$

During inference, the neural network is repeatedly applied until the speech activity  $y_k$  for the last estimated speaker approaches zero. Kinoshita et al. [174] proposed a different approach that combines EEND and speaker clustering. In their method, a neural network is trained to generate speaker embeddings and the speech activity probability. Speaker clustering constrained by the estimated speech activity by EEND is applied to align the estimated speakers among the different processing blocks.

There are also a few recent trials to extend the EEND for on-line processing. Xue et al. [172] proposed a method using a speaker tracing buffer to better align the speaker labels of adjacent processing blocks. Han et al. [173] proposed a block on-line version of EEND-EDA [170] by carrying the hidden state of the LSTM-encoder to generate the attractors block by block.

#### 4. Speaker Diarization in the Context of ASR

From a conventional perspective, speaker diarization is considered a pre-processing step for ASR. In the traditional system structures for speaker diarization, presented in Fig. 1, speech inputs are processed sequentially across the diarization components without considering the ASR performance, which is usually measured using the word error rate (WER). WER is the number of misrecognized words (substitution error, insertion error, and deletion error) divided by the number of reference words. One issue is that the tight boundaries of speech segments as the outcomes of speaker diarization have a high chance of causing unexpected word truncation or deletion errors in ASR decoding. In this section we discuss how the speaker diarization systems have been developed in the context of ASR, not only resulting in better WER by preventing speaker diarization from affecting the ASR performance, but also benefiting from ASR artifacts to enhance diarization performance. More recently, there have been a few pioneering proposals made for the joint modeling of speaker diarization and ASR, which will also be introduced in this section.

##### 4.1. Early Works

The lexical information from the ASR output has been employed for the speaker diarization system in a few different ways. First, the earliest approach was the RT03 evaluation [1] which used word boundary information for the purpose of segmentation. In [1], a general ASR system for broadcast news data was built, in which the basic components are segmentation, speaker clustering, speaker adaptation and system combination after ASR decoding from the two sub-systems with the different adaptation methods. To understand the impact of the word boundary information, they used the ASR outputs to replace the segmentation part and compared the diarization performances of each system. In addition, the ASR result was used to refine SAD in IBM's submission [190] for RT07 evaluation. The system that appeared in [190] incorporates word alignments from the speaker independent ASR module and refines the SAD result to reduce false alarms so that the speaker diarization system can have better clustering quality. The segmentation system in [73] also takes advantage of word alignments from ASR.



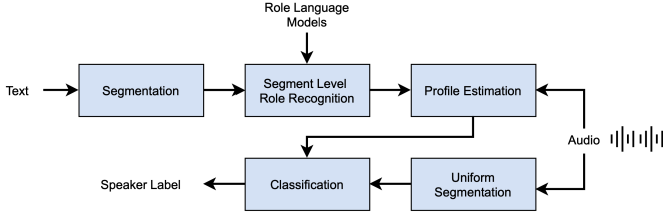


Fig. 14: Integration of lexical information and acoustic information.

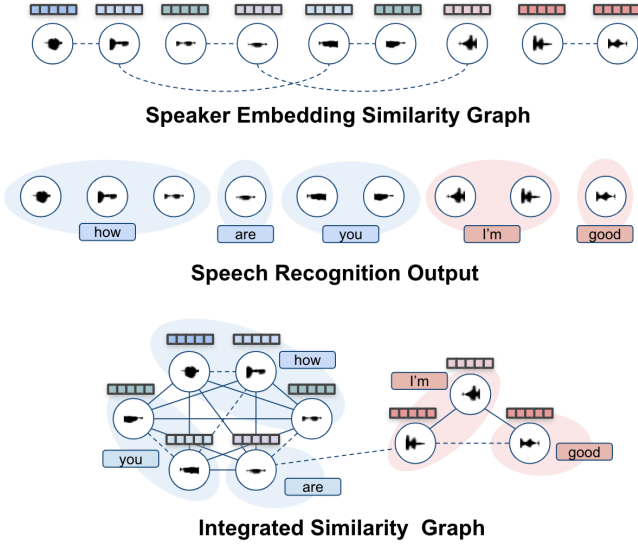


Fig. 15: Integration of lexical information and acoustic information.

The authors in [73] focused on the word-breakage problem, in which the words from the ASR output are truncated by segmentation results since the segmentation results and decoded word sequences are not aligned. Therefore, word-breakage ratio was proposed to measure the rate of change points detected inside intervals corresponding to words. The DER and word-breakage ratio were used to measure the influence of the word truncation problem. While the aforementioned early works of speaker diarization systems that leverage the ASR output focus on the word alignment information to refine the SAD or segmentation result, the speaker diarization system in [191] created a dictionary for the phrases commonly appearing in broadcast news. The phrases in this dictionary provide the identity of who is speaking, who will speak and who spoke in the broadcast news scenario. For example, “This is [name]” indicates who was the speaker of the broadcast news section. Although the early studies on speaker diarization did not fully leverage the lexical information to drastically improve the DER, the idea of integrating the information from ASR output has been adopted by many studies to refine or improve the speaker diarization output.

#### 4.2. Using lexical information from ASR

The more recent speaker diarization systems that take advantage of the ASR transcript have employed a DNN model to capture the linguistic pattern in the given ASR output to enhance

the speaker diarization result. The authors in [192] proposed a way of using the linguistic information for the speaker diarization task where participants have distinct roles that are known to the speaker diarization system. Fig. 14 shows the diagram of the speaker diarization system discussed in [192]. In this system, a neural text-based speaker change detector and a text-based role recognizer are employed. By using both linguistic and acoustic information, DER was significantly improved compared with the acoustic only system.

Lexical information from the ASR output was also used for speaker segmentation [74] by employing a sequence-to-sequence model that outputs speaker turn tokens. Based on the estimated speaker turn, the input utterance is segmented accordingly. The experimental results in [74] indicate that using both acoustic and lexical information can be exploited and an extra advantage can be obtained owing to the word boundaries we get from the ASR output.

[193] presented follow-up research within the above thread. Unlike the system in [74], the lexical information from the ASR module was integrated with the speech segment clustering process by employing an integrated adjacency matrix. The adjacency matrix is obtained from the max operation between the acoustic information created from affinities among audio segments and lexical information matrix created by segmenting the word sequence into word chunks that are likely to be spoken by the same speaker. Fig. 15 presents a diagram that explains how lexical information is integrated in an affinity matrix with acoustic information. The integrated adjacency matrix leads to an improved speaker diarization performance for the CALL-HOME American English dataset.

#### 4.3. Joint ASR and Speaker Diarization with Deep Learning

Motivated by the recent success of deep learning and end-to-end modeling, several models have been proposed to jointly perform ASR and speaker diarization. As with the previous section, the ASR results contain a strong cue to improve speaker diarization. On the other hand, speaker diarization results can be used to improve the accuracy of ASR, for example, by adapting the ASR model toward each estimated speaker. Joint modeling can leverage such inter-dependency to improve both ASR and speaker diarization. In the evaluation, a WER metric that counts word hypotheses with speaker-attribution errors as misrecognized words, such as speaker-attributed WER [194] or concatenated minimum-permutation WER (cpWER) [83], is often used. ASR-specific metrics (e.g., speaker-agnostic WER) or diarization-specific metrics (e.g., DER) are also used complementarily.

The first approach is the introduction of a speaker tag in the transcription of end-to-end ASR models (Fig. 16). Shafey et al. [67] proposed to insert a speaker role tag (e.g., <doctor> and <patient>) into the output of a recurrent neural network-transducer (RNN-T)-based ASR system. This method was evaluated by using doctor-patient conversation, and a significant reduction in WDER was reported with a marginal degradation of WER. Similarly, Mao et al. [68] proposed the insertion of a speaker identity tag into the output of an attention-based encoder-decoder ASR system, and showed an improvement of

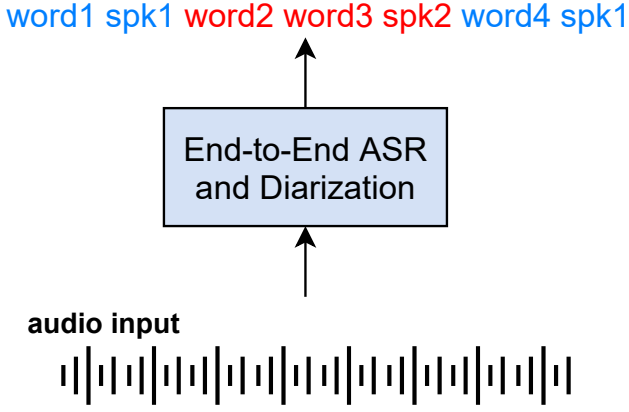


Fig. 16: Joint ASR and diarization by inserting a speaker tag in the transcription.

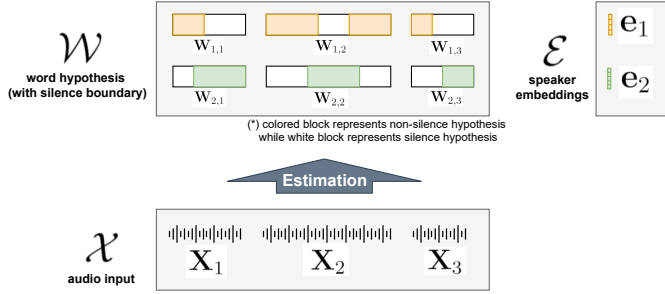


Fig. 17: Joint decoding framework for ASR and speaker diarization.

DER especially when the oracle utterance boundaries were not given. The works by Shafey et al. and Mao et al. showed that the insertion of speaker tags is a simple and promising way to jointly perform ASR and speaker diarization. On the other hand, the speaker roles or speaker identity tags need to be determined and fixed during training. Thus, it is difficult to cope with an arbitrary number of speakers using this approach.

A second approach is a MAP-based joint decoding framework. Kanda et al. [69] formulated the joint decoding of ASR and speaker diarization as follows (see also Fig. 17). Assume that a sequence of observations is represented by  $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_U\}$ , where  $U$  denotes the number of segments (e.g., generated by applying VAD on a long audio) and  $\mathbf{X}_u$  denotes the acoustic feature sequence of the  $u$ -th segment. Further assume that word hypotheses with time boundary information are represented by  $\mathcal{W} = \{\mathbf{W}_1, \dots, \mathbf{W}_U\}$  where  $\mathbf{W}_u$  is the speech recognition hypothesis corresponding to the segment  $u$ . Here,  $\mathbf{W}_u = (\mathbf{W}_{1,u}, \dots, \mathbf{W}_{K,u})$  contains all the speakers' hypotheses in the segment  $u$  where  $K$  denotes the number of speakers, and  $\mathbf{W}_{k,u}$  represents the speech recognition hypothesis of the  $k$ -th speaker of the segment  $u$ . Finally, a tuple of speaker embeddings  $\mathcal{E} = (\mathbf{e}_1, \dots, \mathbf{e}_K)$ , where  $\mathbf{e}_j \in \mathbb{R}^d$  is the  $d$ -dimensional speaker embedding of the  $k$ -th speaker, is also assumed. With all these notations, the joint decoding framework of multi-speaker ASR and diarization can be formulated as a problem

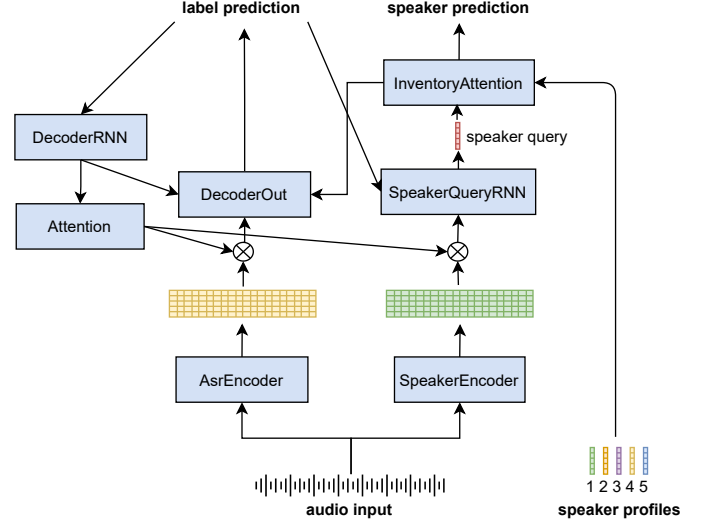


Fig. 18: End-to-end speaker-attributed ASR

to find most likely  $\hat{\mathcal{W}}$  as follows:

$$\hat{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmax}} P(\mathcal{W}|\mathcal{X}) \quad (33)$$

$$= \underset{\mathcal{W}}{\operatorname{argmax}} \left\{ \sum_{\mathcal{E}} P(\mathcal{W}, \mathcal{E}|\mathcal{X}) \right\} \quad (34)$$

$$\approx \underset{\mathcal{W}}{\operatorname{argmax}} \left\{ \max_{\mathcal{E}} P(\mathcal{W}, \mathcal{E}|\mathcal{X}) \right\}, \quad (35)$$

where the Viterbi approximation is applied to obtain the final equation. This maximization problem is further decomposed into two iterative problems as follows:

$$\hat{\mathcal{W}}^{(i)} = \underset{\mathcal{W}}{\operatorname{argmax}} P(\mathcal{W}|\hat{\mathcal{E}}^{(i-1)}, \mathcal{X}), \quad (36)$$

$$\hat{\mathcal{E}}^{(i)} = \underset{\mathcal{E}}{\operatorname{argmax}} P(\mathcal{E}|\hat{\mathcal{W}}^{(i)}, \mathcal{X}), \quad (37)$$

where  $i$  is the iteration index of the procedure. In [69], Eq. (36) is modeled by the target speaker ASR [195, 196, 197, 61] and Eq. (37) is modeled by the overlap-aware speaker embedding estimation. This method obtains a speaker-attributed WER similar to that of the target-speaker ASR with oracle speaker embeddings for two-speaker conversation data of the Corpus of Spontaneous Japanese [189]. On the other hand, it requires an iterative application of the target-speaker ASR and speaker embedding extraction, which makes it challenging to apply the method in online mode.

As a third line of approaches, end-to-end speaker-attributed ASR (SA-ASR) model was recently proposed to jointly perform speaker counting, multi-talker ASR, and speaker identification [175, 176]. Different from the first two approaches, the end-to-end SA-ASR model takes the additional input of speaker profiles and identifies the index of speaker profiles based on the attention mechanism (Fig. 18). Thanks to the attention mechanism for the speaker identification and multi-talker ASR capability based on serialized output training [198], there is no limitation in the number of speakers that the model can cope with. In case relevant speaker profiles are supplied in the inference,

the end-to-end SA-ASR model can automatically transcribe the utterance while identifying the speaker of each utterance based on the supplied speaker profiles. On the other hand, in case the relevant speaker profiles cannot be used prior to the inference, the end-to-end SA-ASR model can still be applied using dummy profiles, and speaker clustering on the internal speaker embeddings of the end-to-end SA-ASR model (“speaker query” in Fig. 18) is used to diarize the speaker [70]. The end-to-end SA-ASR model was evaluated by using the LibriCSS dataset [81], and exhibited significantly better cpWER than the combination of multitalker ASR and speaker diarization [199].

## 5. Diarization Evaluation Series and Datasets

This section describes the evaluation series and the commonly used datasets for speaker diarization evaluations. The summary of the most commonly used datasets that include English is shown in Table 3.

- **CALLHOME: NIST SRE 2000 (LDC2001S97)**  
NIST SRE 2000 (Disk-8), often referred to as the CALLHOME dataset, is the most widely used dataset for speaker diarization in recent papers. This dataset contains 500 sessions of multilingual telephonic speech. Each session has two to seven speakers with two dominant speakers in each conversation.
- **AMI Corpus**  
The AMI database [200] includes 100 h of meeting recordings from multiple sites in 171 meeting sessions. The AMI database provides an audio source recorded using lapel microphones that are separately recorded and amplified for each speaker. Another audio source is recorded using omnidirectional microphone arrays mounted on the table while meeting. The AMI database is a suitable dataset for the evaluation of speaker diarization systems integrated with the ASR module since AMI provides forced alignment data that contains word and phoneme level timings along with the transcript and speaker label. Each meeting session has three to five speakers.
- **ICSI Meeting Corpus**  
The ICSI meeting corpus [201] contains 75 meeting corpus with four meeting types. The ICSI meeting corpus provides word level timing along with the transcript and speaker label. The audio source is recorded using close-talking individual microphone and six tabletop microphones to provide speaker-specific channel and multi-channel recording. Each meeting has 3 to 10 participants.
- **CHiME-5/6 challenge and its dataset**  
The CHiME-5 challenge [202] and CHiME-6 challenge [83] were designed as series of ASR competitions for the daily conversation of multiple speakers. The dataset was provided at the CHiME-5 challenge, and it contains 50 h of multiparty real conversations in the everyday home environment. It contains speaker labels, segmentation, and corresponding transcriptions. The audio source

is recorded using six four-channel microphone arrays located in the kitchen and dining/living rooms in a house and also binaural microphones worn by participants. The number of participants is fixed at four. While the oracle diarization results were allowed to be used for the ASR task in the CHiME-5 challenge, CHiME-6 challenge track 2 requires the performance of both ASR and diarization. The primary evaluation metric for such a track was cpWER, which counts both the speaker-attributed errors and word recognition errors in the WER calculation. DER and JER were also evaluated as secondary metrics without “score collar” and with overlapped regions. The CHiME-5/6 corpus was also used as one track in the DIHARD 2 challenge.

- **VoxSRC Challenge and VoxConverse corpus**  
The VoxCeleb Speaker Recognition Challenge (VoxSRC) is the recent evaluation series for speaker recognition systems [203, 97]. The goal of VoxSRC is to test how well the current technology can cope with the speech “in the wild”. This evaluation series initially started with a pure speaker verification task [203], and the diarization task was added as track 4 at the latest evaluation at the VoxCeleb Speaker Recognition Challenge 2020 (VoxSRC-20) [97]. The VoxConverse dataset [204] was used for the speaker diarization task with DER as the primary metric, and JER as the secondary metric. The VoxConverse dataset contains 74 h of human conversation extracted from YouTube videos. The dataset is divided into development set (20.3 h, 216 recordings), and test set (53.5 h, 310 recordings). The number of speakers in each recording has a wide range of variety from 1 speaker to 21 speakers. The audio includes various types of noises such as background music, laughter, etc. It also contains a significant portion of overlapping speech from 0% to 30.1% depending on the recording. While the dataset contains the visual information as well as audio, as of June 2021, only the audio of the development set was released under a Creative Commons Attribution 4.0 International License for research purposes. The audio of the evaluation set was used as a blind test set.
- **LibriCSS**  
The LibriCSS corpus [81] contains 10 h of multichannel recordings and was designed for the research of speech separation, speech recognition, and speaker diarization. It was created by playing back the audio in the LibriSpeech corpus [205] in a real meeting room, and recorded using a 7-channel microphone array. It consists of 10 sessions, each of which is further decomposed to six 10-min mini-sessions. Each mini-session was made by audio of eight speakers and designed to have different overlap ratios from 0% to 40%. To facilitate the research, the baseline system for speech separation and ASR [81] and the baseline system that integrates speech separation, speaker diarization and ASR [82] have been developed and released.

Table 3: Diarization Evaluation Datasets

	Language	Size (hr)	Style	# Spkr.
CALLHOME	Multilingual	20	Conversation	2–7
AMI	English	100	Meeting	3–5
ICSI meeting	English	72	Meeting	3–10
CHiME-5/6	English	50	Conversation	4
VoxConverse	Multilingual <sup>†</sup>	74	YouTube video	1–21
LibriCSS	English	10	Read speech	8
DH I Tr.1,2	Multilingual	19(dev), 21(eval)	Miscellaneous	1–7
DH II Tr.1,2	Multilingual	24(dev), 22(eval)	Miscellaneous	1–8
DH II Tr.3,4	Multilingual	262(dev), 31(eval)	Miscellaneous	4
DH III Tr.1,2	Multilingual	34(dev), 33(eval)	Miscellaneous	1–7

<sup>†</sup> Most of the contents are English while there are few non-English contents.

#### • DIHARD Challenge and its dataset

DIHARD evaluation [206, 79] focuses on the performance gap of state-of-the-art diarization systems on challenging domains. The first DIHARD challenge, DIHARD 1, started with track 1 (oracle SAD) and track 2 (system SAD). The evaluation data was a collection of various corpus. It includes very challenging datasets such as clinical interviews, web videos, and speech in the wild (e.g., recordings in restaurants), as well as relatively less challenging datasets, such as CTS and audio books to diversify the domains. DIHARD 2 additionally included multichannel speaker diarization task in track 3 (oracle SAD) and track 4 (system SAD) using the recordings from the CHiME-5 corpus [202]. In the latest DIHARD challenge, DIHARD 3, the CTS dataset was added, whereas multichannel tracks 3 and 4 were excluded. The DIHARD challenge employs DER and JER for the evaluation metric without “score collar” and with overlap region.

#### • Rich Transcription Evaluation Series

The RT Evaluation [13] is the pioneering evaluation series of initiating deeper investigation on speaker diarization in relation to ASR. The main purpose of this effort was to create ASR technologies that would produce transcriptions with descriptive metadata, like who said when, where speaker diarization plays in. Thus, the main tasks in the evaluation were ASR and speaker diarization. The domains of the data of interest were broadcast news, CTS and meeting recordings with multiple participants. Throughout the period of 2002-2009, the RT evaluation series promoted and gauged advances in speaker diarization and ASR technology. The evaluations among this period are named as RT evaluations (RT-02, RT-03S, RT-03F, and RT-05F) and RT Meeting Recognition (RT-06S, RT-07S and RT-09). These evaluations and their datasets include speaker diarization evaluation as a part of automatic metadata extraction (MDE).

#### • Other datasets

There are also several corpora that have been used for the diarization research but not covered in the list above. The Corpus of Spontaneous Japanese [189] contains about 12

h of two-speaker dialogue recorded using headset microphones. AISHELL-4 [207] is a relatively new Mandarin Chinese dataset containing 118 h of four to eight speakers in a conference scenario. It is recorded by 8-ch circular microphone array as well as headset microphones for each participant. The ESTER-1 [208] and ESTER-2 [209] evaluation campaign datasets are a set of French recordings designed for three task category: Segmentation (S), Transcription (T), and Information Extraction (E). In the ESTER-1 and ESTER-2 evaluation campaign, speaker diarization was evaluated as one of the core tasks among other tasks including speaker tracking, sound event tracking, and transcriptions. The datasets for ESTER-1 and ESTER-2 include 100 h and 150 h of manually transcribed French radio broadcast news, respectively. ETAPE [210] is also a French speech processing evaluation dataset that contains 36 h of TV and radio shows with both prepared and spontaneous speech. Unlike the ESTER evaluation series, ETAPE targets cross-show speaker diarization.

## 6. Applications

### 6.1. Meeting Transcription

The goal of meeting transcription is to automatically generate speaker-attributed transcripts during real-life meetings based on their audio and optional video recordings. Accurate meeting transcription is among the processing steps in a pipeline for several tasks, such as, summarization and topic extraction. Similarly, the same transcription system can be used in other domains such as healthcare [211].

Although this task was introduced by NIST in the RT Evaluation series back in 2003 [194, 201, 212], the initial systems had very poor performance, and consequently commercialization of the technology was not possible. However, recent advances in the areas of speech recognition [213, 214], far-field speech processing [215, 216, 217], speaker ID and diarization [218, 43, 78], have greatly improved the speaker-attributed transcription accuracy, enabling such commercialization. Bimodal processing combining cameras with microphone arrays has further improved the overall performance [219, 220].



The variety of application scenarios, customer needs, and business scope, different constraints may be imposed on meeting transcription systems. For example, it is most often required to provide the resulting transcriptions in low latency, making the diarization and recognition even more challenging. However, the architecture of the transcription system can substantially improve the overall performance, e.g., by using microphone arrays of known geometry as the input device. Also, in the case where the expected meeting attendees are known beforehand, the transcription system can further improve speaker attribution, all while providing the exact name of the speaker, instead of randomly generated discrete speaker labels.

Two different scenarios in this space are presented: first, a fixed-geometry microphone array combined with a fish-eye camera system. Second, an ad-hoc geometry microphone array system without a camera. In both scenarios, a “non-binding” list of participants and their corresponding speaker profiles are considered to be known. In particular, the transcription system has access to the invitees’ names and profiles; however, the actual attendees may not accurately match those invited. As such, there is an option to include “unannounced” participants. In addition, some of the invitees may not have profiles. In both scenarios, there is a constraint of low-latency transcriptions, where initial results need to be shown with low latency. The finalized results can be updated later in an offline mode. Some of the technical challenges to overcome are [221]:

1. Although ASR on overlapping speech is one of the main challenges in meeting transcription, limited progress has been made over the years. Numerous multichannel speech separation methods have been proposed based on independent component analysis (ICA) or spatial clustering [222, 223, 224, 225, 226, 227], but their application to a meeting setup had limited success. In addition, neural network-based separation methods such as permutation invariant training (PIT) [47] or deep clustering (DC) [46] cannot adequately address reverberation and background noise [228].
2. Flexible framework: It is desirable that the transcription system is capable of processing all the available information, such as the multichannel audio and visual cues. The system needs to process a dynamically changing number of audio channels without loss of performance. As such, the architecture needs to be modular enough to encompass the different settings.
3. The speaker-attributed ASR of natural meetings requires online or streaming ASR, audio pre-processing such as dereverberation, and accurate diarization and speaker identification. These multiple processing steps are usually optimized separately and thus, the overall pipeline is most frequently inefficient.
4. The use of multiple, unsynchronized audio streams, e.g., audio capturing using mobile devices, adds complexity to the meeting setup and processing. In return, we gain a potentially better spatial coverage since the devices are usually distributed around the room and near the speakers. As

part of the application scenario, the meeting participants bring their personal devices, which can be repurposed to improve the overall quality of meeting transcription quality. On the other hand, while there are several pioneering studies [229], it is unclear what the best strategies are for consolidating multiple asynchronous audio streams and to what extent they work for natural meetings in online and offline setups.

Based on these considerations, an architecture of a meeting transcription system with asynchronous distant microphones has been proposed in [184]. In this work, various fusion strategies have been investigated: from early fusion beamforming of the audio signals, to mid-fusion combination of senones per channel, to late fusion combination of the diarization and ASR results [161]. The resulting system performance was benchmarked on real-world meeting recordings against fixed geometry systems. As aforementioned, the requirement of speaker-attributed transcriptions with low latency was also adhered to. In addition to the end-to-end system analysis, the paper [184] proposed the idea of “leave-one-out beamforming” in the asynchronous multi-microphone setup, enriching the “diversity” of the resulting signals, as proposed in [230]. Finally, it is described how an online, incremental version of recognizer output voting error reduction (ROVER) [157] can process both the ASR and diarization outputs, enhancing the overall speaker-attributed ASR performance.

## 6.2. Conversational Interaction Analysis and Behavioral Modeling

Speech and spoken language are central to conversational interactions. They carry crucial information about a speaker’s intent, emotions, identity, age and other individual and interpersonal traits and state variables including health state. Computational advances are increasingly allowing access to such rich information [231, 232]. For example, knowing how much, and how, a child speaks in an interaction contains critical information about the child’s developmental state, and offers clues to clinicians in diagnosing disorders such as autism [233]. Such analyses are made possible by capturing and processing the audio recordings of the interactions, which often involve two or more people. An important foundational step is the identification and association of the speech portions belonging to specific individuals involved in the conversation. The technologies providing these capabilities are SAD and speaker diarization. Speech portions segmented with speaker-specific information provided by speaker diarization, by itself without any explicit lexical transcription, can offer important information to domain experts who can take advantage of speaker diarization results for quantitative turn-taking analysis.

A domain that is the most relevant in such analyses of spoken conversational interactions relates to behavioral signal processing (BSP) [234, 231], which refers to the technology and algorithms for modeling and understanding human communicative, affective and social behaviors. For example, these may include analyzing how positive or negative a person is, how empathic

an individual is toward another, what the behavior patterns reveal about the relationship status, and the health condition of an individual [232]. BSP involves addressing all the complexities of spontaneous interactions in conversations with additional challenges involved in handling and understanding emotional, social and interpersonal behavioral dynamics revealed through verbal and nonverbal cues of the interaction participants. Therefore, the knowledge of speaker specific vocal information plays a significant role in BSP, requiring highly accurate speaker diarization performance. For example, the speaker diarization module is employed as a pre-processing module for analyzing psychotherapy mechanisms and quality [235] and suicide risk assessment [236].

Another popular application of speaker diarization for conversation interaction analysis is the medical doctor-patient interactions. In the system described in [237], the nature of memory problems of a patient is detected from the conversations between neurologists and patients. The speech and language features extracted from the ASR transcripts combined with the speaker diarization results are used to predict the type of disorder. An automated assistant system for medical domain transcription is proposed in [238], which includes the speaker diarization module, ASR module and natural language generation module. The automated assistant module accepts the audio clip and outputs grammatically correct sentences describing the topic of the conversation, subject and subject's symptom.

### 6.3. Audio Indexing

Content-based audio indexing is a well known application domain for speaker diarization. It can provide meta-information such as the content or data type of a given audio data to make information retrieval efficient since search query by machines would be limited by such metadata. The more diverse information was available, the better efficiency could be achieved in retrieving audio contents from a database.

One useful piece of information for the audio indexing would be ASR transcripts to understand the content of speech portions in the audio data. Speaker diarization can augment those transcripts in terms of "who spoke when", which was the main purpose of the RT evaluation series [13], as discussed in Sections 4.1 and 5.3. The aggregated spoken utterances from speakers by a speaker diarization system also enable per-speaker summary or keyword list-up, which can be used for other query values to retrieve relevant contents from the database. In [239], we can get a view of how speaker diarization outputs can be linked for information searching in consumer-facing applications.

### 6.4. Conversational AI

Thanks to the advances of ASR technology, the applications of ASR have evolved from simple voice command recognition systems to conversational AI systems. Conversational AI systems, as opposed to voice command recognition systems, have features that are lacking in voice command recognition systems. The fundamental idea of conversational AI is to build a machine that humans can talk to and interact with. In this sense, focusing on an interested speaker in a multiparty setting is one

of the most important features of conversational AI. Moreover, speaker diarization becomes an essential feature for conversational AI. For example, conversational AI equipped in a car can pay attention to a specific speaker that is demanding a piece of information from the navigation system by applying speaker diarization along with ASR.

Smart speakers and voice assistants are the most popular products in which speaker diarization plays a significant role for conversational AI. Since the response time and online processing are the crucial factors in real-life settings, the demand for end-to-end speaker diarization systems integrated into the ASR pipeline is growing. The performance of incremental (online) ASR and speaker diarization of the commercial ASR services are evaluated and compared in [240]. It is expected that the real-time and low latency aspect of speaker diarization will be more emphasized in the speaker diarization systems in the future since the performance of online diarization and online ASR still have much room for improvement.

## 7. Challenges and the Future of Speaker Diarization

This paper has provided a comprehensive overview of speaker diarization techniques, highlighting the recent development of deep learning-based diarization approaches. In the early days, a speaker diarization system was developed as a pipeline of sub-modules including front-end processing, SAD, segmentation, speaker embedding extraction, clustering, and post-processing, leading to a standalone system without much connection to other components in a given speech application. With the emergence of the deep learning technology, more and more advancements have been made for speaker diarization, from a method that replaces a single module into a deep-learning-based method, to a fully EEND. Furthermore, as the speech recognition technology becomes more accessible, a trend to tightly integrate speaker diarization into the ASR systems has emerged, such as benefiting from the ASR output to improve the accuracy of speaker diarization. Recently, joint modeling for speaker diarization and speech recognition is investigated in an attempt to enhance the overall performance of speaker diarization. Thanks to these great achievements, speaker diarization systems have already been used in many applications, including meeting transcription, conversational interaction analysis, audio indexing, and conversational AI systems.

As we have seen, tremendous progress has been made for speaker diarization systems. Nevertheless, there is still much room for improvement. As the final remark, we conclude this paper by listing the remaining challenges for speaker diarization toward future research and development.

*Online processing of speaker diarization.* Most speaker diarization methods assume that an entire recording can be observed to execute speaker diarization. However, numerous applications such as meeting transcription systems or smart agents require very short latency for assigning the speaker. While there have been several attempts to make online speaker diarization

system both for clustering-based systems (e.g., [218]) and neural network-based diarization systems (e.g., [43, 172, 173]), it still remains as a challenging problem.

*Domain mismatch.* A model that is trained on a data in a specific domain often works poorly on data in another domain. For example, it is experimentally known that the EEND model tends to overfit to the distribution of the speaker overlaps of the training data [44]. Such a domain mismatch issue is universal for any training-based method. Given the growing interest for trainable speaker diarization systems, it will become more important to evaluate the ability for handling the variety of inputs. The international evaluation efforts for speaker diarization such as the DIHARD challenge [206, 79, 241] or VoxSRC [203, 97] also have great importance in that direction.

*Speaker overlap.* Overlap of multitalker speech is the inevitable nature of conversation. For example, an average of 12% to 15% of speaker overlap was observed in meeting recordings [242, 94], and it could even increase in daily conversations [243, 202, 83]. Nevertheless, many traditional speaker diarization systems, especially clustering-based systems, have only focused on non-overlapping regions and even the overlapping regions are excluded in the evaluation metric [244]. While the topic has been studied for long years (e.g. early works [245, 246]), there is a growing interest for handling the speaker overlaps toward better speaker diarization, including the application of speech separation [96], post-processing [247, 168], and joint modeling of speech separation and speaker diarization [66, 175].

*Integration with ASR.* Many applications require ASR results along with speaker diarization results. In the modular combination of speaker diarization and ASR, some systems locate a speaker diarization system before ASR [85] while some systems locate a diarization system after ASR [221, 193]. Both types of systems showed a strong performance for a specific task, and determining the best kind of system architecture for the speaker diarization and ASR tasks is still an open problem [82]. Furthermore, there is another line of research to jointly perform speaker diarization and ASR [67, 68, 69, 175], which was introduced in Section 4. The joint modeling approach could leverage the inter-dependency between speaker diarization and ASR to better perform both tasks. However, it has not yet been fully investigated whether such joint frameworks perform better than the well-tuned modular systems. Overall, the integration of speaker diarization and ASR is one of the hottest topics that is still being investigated by many researchers.

*Audiovisual modeling.* Visual information contains a strong clue for the identification of speakers. For example, the video captured by a fish-eye camera was used to improve the accuracy of speaker diarization in a meeting transcription task [221]. The visual information was also used to significantly improve the accuracy of speaker diarization for speaker diarization on YouTube video [204]. While these studies showed the effectiveness of visual information, the audiovisual speaker diarization has yet been rarely investigated compared with audio-only

speaker diarization, and there will be many rooms for improvement.

## References

- [1] S. E. Tranter, K. Yu, D. A. Reynolds, G. Evermann, D. Y. Kim, P. C. Woodland, An investigation into the interactions between speaker diarisation systems and automatic speech transcription, CUED/F-INFENG/TR-464 (2003).
- [2] S. E. Tranter, D. A. Reynolds, An overview of automatic speaker diarization systems, IEEE Transactions on Audio, Speech, and Language Processing 14 (2006) 1557–1565.
- [3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals, Speaker diarization: A review of recent research, IEEE Transactions on Audio, Speech, and Language Processing 20 (2012) 356–370.
- [4] H. Gish, M. . Siu, R. Rohlicek, Segregation of speakers for speech recognition and speaker identification, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1991, pp. 873–876.
- [5] M.-H. Siu, Y. George, H. Gish, An unsupervised, sequential learning algorithm for segmentation for speech waveforms with multiple speakers, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1992, pp. 189–192.
- [6] J. R. Rohlicek, D. Ayuso, M. Bates, R. Bobrow, A. Boulanger, H. Gish, P. Jeanrenaud, M. Meteer, M. Siu, Gisting conversational speech, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1992, pp. 113–116.
- [7] U. Jain, M. A. Siegler, S.-J. Doh, E. Gouvea, J. Huerta, P. J. Moreno, B. Raj, R. M. Stern, Recognition of continuous broadcast news with multiple unknown speakers and environments, in: Proceedings of ARPA Spoken Language Technology Workshop, 1996, pp. 61–66.
- [8] M. Padmanabhan, L. R. Bahl, D. Nahamoo, M. A. Picheny, Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1996, pp. 701–704.
- [9] J.-L. Gauvain, L. Lamel, G. Adda, Partitioning and transcription of broadcast news data, in: Proceedings of the International Conference on Spoken Language Processing, 1998, pp. 1335–1338.
- [10] D. Liu, F. Kubala, Fast speaker change detection for broadcast news transcription and indexing, in: Proceedings of the International Conference on Spoken Language Processing, 1999, pp. 1031–1034.
- [11] S. S. Chen, P. S. Gopalakrishnan, Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion, in: Tech. Rep., IBM T. J. Watson Research Center, 1998, pp. 127–132.
- [12] AMI, AMI Consortium. <http://www.amiproject.org/index.html>.
- [13] NIST, Rich Transcription Evaluation. <https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation>.
- [14] J. Ajmera, C. Wooters, A robust speaker clustering algorithm, in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, 2003, pp. 411–416.
- [15] S. E. Tranter, D. A. Reynolds, Speaker diarisation for broadcast news, in: Odyssey, 2004, pp. 337–344.
- [16] D. A. Reynolds, P. Torres-Carrasquillo, Approaches and applications of audio diarization, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2005, pp. 953–956.
- [17] X. Zhu, C. Barras, S. Meignier, J.-L. Gauvain, Combining speaker identification and BIC for speaker diarization, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2005, pp. 2441–2444.
- [18] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, L. Besacier, Step-by-step and integrated approaches in broadcast news speaker diarization, Computer, Speech & Language 20 (2006) 303–330.
- [19] A. E. Rosenberg, A. Gorin, Z. Liu, P. Parthasarathy, Unsupervised speaker segmentation of telephone conversations, in: Proceedings of the International Conference on Spoken Language Processing, 2002, pp. 565–568.
- [20] D. Liu, F. Kubala, A cross-channel modeling approach for automatic segmentation of conversational telephone speech, in: Proceedings of

- IEEE Workshop on Automatic Speech Recognition and Understanding, 2003, pp. 333–338.
- [21] S. E. Tranter, K. Yu, G. Evermann, P. C. Woodland, Generating and evaluating for automatic speech recognition of conversational telephone speech, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2004, pp. 753–756.
  - [22] P. Kenny, D. Reynolds, F. Castaldo, Diarization of telephone conversations using factor analysis, *IEEE Journal of Selected Topics in Signal Processing* 4 (2010) 1059–1070.
  - [23] J. Ajmera, G. Lathoud, L. McCowan, Clustering and segmenting speakers and their locations in meetings, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2004, pp. 605–608.
  - [24] Q. Jin, K. Laskowski, T. Schultz, A. Waibel, Speaker segmentation and clustering in meetings, in: Proceedings of the International Conference on Spoken Language Processing, 2004, pp. 597–600.
  - [25] X. Anguera, C. Wooters, J. Hernando, Purity algorithms for speaker diarization of meetings data, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume I, 2006, pp. 1025–1028.
  - [26] D. A. V. Leeuwen, M. Konecny, Progress in the AMIDA speaker diarization system for meeting data, in: Proceedings of International Evaluation Workshops CLEAR 2007 and RT 2007, 2007, pp. 475–483.
  - [27] D. Vijayasenan, F. Valente, H. Bourlard, An information theoretic approach to speaker diarization of meeting data, *IEEE Transactions on Audio, Speech, and Language Processing* 17 (2009) 1382–1393.
  - [28] X. Anguera, C. Wooters, J. Hernando, Acoustic beamforming for speaker diarization of meetings, *IEEE Transactions on Audio, Speech, and Language Processing* 15 (2007) 2011–2023.
  - [29] F. Valente, P. Motlicek, D. Vijayasenan, Variational Bayesian speaker diarization of meeting recordings, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 4954–4957.
  - [30] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing* 19 (2011).
  - [31] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair, Stream-based speaker segmentation using speaker factors and eigenvoices, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 4133–4136.
  - [32] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, J. Glass, Exploiting intra-conversation variability for speaker diarization, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2011.
  - [33] S. Shum, N. Dehak, J. Glass, On the use of spectral and iterative methods for speaker diarization, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2012, pp. 482–485.
  - [34] S. H. Shum, N. Dehak, R. Dehak, J. R. Glass, Unsupervised methods for speaker diarization: An integrated and iterative approach, *IEEE Transactions on Audio, Speech, and Language Processing* 21 (May 2013) 2015–2028.
  - [35] M. Senoussaoui, P. Kenny, T. Stafylakis, P. Dumouchel, A study of the cosine distance-based mean shift for telephone speech diarization, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22 (2013) 217–227.
  - [36] G. Sell, D. Garcia-Romero, Speaker diarization with plda i-vector scoring and unsupervised calibration, in: Proceedings of IEEE Spoken Language Technology Workshop, IEEE, 2014, pp. 413–417.
  - [37] E. Variani, X. Lei, E. McDermott, I. L. Moreno, J. G-Dominguez, Deep neural networks for small footprint text-dependent speaker verification, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, pp. 4052–4056.
  - [38] G. Heigold, I. Moreno, S. Bengio, N. Shazeer, End-to-end text-dependent speaker verification, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, pp. 5115–5119.
  - [39] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, I. L. Moreno, Speaker diarization with LSTM, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5239–5243.
  - [40] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5329–5333.
  - [41] G. Dupuy, M. Rouvier, S. Meignier, Y. Esteve, i-Vectors and ILP clustering adapted to cross-show speaker diarization, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2012, pp. 2174–2177.
  - [42] S. H. Shum, N. Dehak, R. Dehak, J. R. Glass, Unsupervised methods for speaker diarization: An integrated and iterative approach, *IEEE Transactions on Audio, Speech, and Language Processing* 21 (2013).
  - [43] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, C. Wang, Fully supervised speaker diarization, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6301–6305.
  - [44] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with permutation-free objectives, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 4300–4304.
  - [45] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with self-attention, in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE, 2019, pp. 296–303.
  - [46] J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, Deep clustering: Discriminative embeddings for segmentation and separation, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2016, pp. 31–35.
  - [47] M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (2017) 1901–1913.
  - [48] Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27 (2019) 1256–1266.
  - [49] E. Variani, X. Lei, E. McDermott, I. L. Moreno, J. Gonzalez-Dominguez, Deep neural networks for small footprint text-dependent speaker verification, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2014, pp. 4052–4056.
  - [50] D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, Deep neural network embeddings for text-independent speaker verification, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2017, pp. 999–1003.
  - [51] T. Drugman, Y. Stylianou, Y. Kida, M. Akamine, Voice activity detection: Merging source and filter-based information, *IEEE Signal Processing Letters* 23 (2015) 252–256.
  - [52] J. Wang, X. Xiao, J. Wu, R. Ramamurthy, F. Rudzicz, M. Brudno, Speaker diarization with session-level speaker embedding refinement using graph neural networks, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 7109–7113.
  - [53] D. Dimitriadis, Enhancements for Audio-only Diarization Systems, *arXiv preprint arXiv:1909.00082* (2019).
  - [54] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko, Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2020, pp. 274–278.
  - [55] M. Diez, L. Burget, P. Matejka, Speaker diarization based on bayesian hmm with eigenvoice priors, in: *Odyssey*, 2018, pp. 147–154.
  - [56] M. Diez, L. Burget, S. Wang, J. Rohdin, J. Cernocký, Bayesian HMM based x-vector clustering for speaker diarization, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 346–350.
  - [57] D. Yu, X. Chang, Y. Qian, Recognizing multi-talker speech with permutation invariant training, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2017, pp. 2456–2460.
  - [58] H. Seki, T. Hori, S. Watanabe, J. Le Roux, J. R. Hershey, A purely end-to-end system for multi-speaker speech recognition, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, volume 1, 2018, pp. 2620–2630.
  - [59] X. Chang, Y. Qian, K. Yu, S. Watanabe, End-to-end monaural multi-



- speaker ASR system without pretraining, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6256–6260.
- [60] N. Kanda, Y. Fujita, S. Horiguchi, R. Ikeshita, K. Nagamatsu, S. Watanabe, Acoustic modeling for distant multi-talker speech recognition with single-and multi-channel branches, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6630–6634.
- [61] N. Kanda, S. Horiguchi, R. Takashima, Y. Fujita, K. Nagamatsu, S. Watanabe, Auxiliary interference speaker loss for target-speaker speech recognition, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 236–240.
- [62] X. Wang, N. Kanda, Y. Gaur, Z. Chen, Z. Meng, T. Yoshioka, Exploring end-to-end multi-channel asr with bias information for meeting transcription, in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, 2021.
- [63] P. Wang, Z. Chen, X. Xiao, Z. Meng, T. Yoshioka, T. Zhou, L. Lu, J. Li, Speech separation using speaker inventory, in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, 2019, pp. 230–236.
- [64] C. Han, Y. Luo, C. Li, T. Zhou, K. Kinoshita, S. Watanabe, M. Delcroix, H. Erdogan, J. R. Hershey, N. Mesgarani, et al., Continuous speech separation using speaker inventory for long multi-talker recording, arXiv preprint arXiv:2012.09727 (2020).
- [65] Z. Huang, S. Watanabe, Y. Fujita, P. García, Y. Shao, D. Povey, S. Khudanpur, Speaker diarization with region proposal network, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 6514–6518.
- [66] T. von Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, R. Haeb-Umbach, All-neural online source separation, counting, and diarization for meeting analysis, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2019, pp. 91–95.
- [67] L. E. Shafey, H. Soltan, I. Shafan, Joint Speech Recognition and Speaker Diarization via Sequence Transduction, in: Proceedings of the Annual Conference of the International Speech Communication Association, ISCA, 2019, pp. 396–400.
- [68] H. H. Mao, S. Li, J. McAuley, G. Cottrell, Speech recognition and multi-speaker diarization of long conversations, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2020, pp. 691–695.
- [69] N. Kanda, S. Horiguchi, Y. Fujita, Y. Xue, K. Nagamatsu, S. Watanabe, Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models, in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, 2019, pp. 31–38.
- [70] N. Kanda, X. Chang, Y. Gaur, X. Wang, Z. Meng, Z. Chen, T. Yoshioka, Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings, in: Proceedings of IEEE Spoken Language Technology Workshop, 2021.
- [71] J. G. Fiscus, J. Ajot, M. Michel, J. S. Garofolo, The rich transcription 2006 spring meeting recognition evaluation, in: Proceedings of International Workshop on Machine Learning and Multimodal Interaction, NIST, 2006, pp. 309–322.
- [72] H. W. Kuhn, The hungarian method for the assignment problem, *Naval research logistics quarterly* 2 (1955) 83–97.
- [73] J. Silovsky, J. Zdansky, J. Nouza, P. Cerva, J. Prazak, Incorporation of the asr output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams, in: International Workshop on Multimedia Signal Processing, IEEE, 2012, pp. 118–123.
- [74] T. J. Park, P. Georgiou, Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 1373–1377.
- [75] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, M. Souden, Speech processing for digital home assistants: Combining signal processing with deep-learning techniques, *IEEE Signal Processing Magazine* 36 (2019) 111–124.
- [76] E. Vincent, T. Virtanen, S. Gannot, Audio source separation and speech enhancement, John Wiley & Sons, 2018.
- [77] D. Wang, J. Chen, Supervised speech separation based on deep learning: An overview, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (2018) 1702–1726.
- [78] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, et al., Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 2808–2812.
- [79] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The second DIHARD diarization challenge: Dataset, task, and baselines, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 978–982.
- [80] M. Diez, F. Landini, L. Burget, J. Rohdin, A. Silnova, K. Zmolíková, O. Novotný, K. Veselý, O. Glembek, O. Plchot, et al., BUT system for DIHARD speech diarization challenge 2018., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 2798–2802.
- [81] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, J. Li, Continuous speech separation: Dataset and analysis, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 7284–7288.
- [82] D. Raj, P. Denirov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, J. R. Hershey, Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis, in: Proceedings of IEEE Spoken Language Technology Workshop, 2021.
- [83] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, et al., CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings, in: 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020), 2020.
- [84] A. Arora, D. Raj, A. S. Subramanian, K. Li, B. Ben-Yair, M. Maciejewski, P. Želasko, P. Garcia, S. Watanabe, S. Khudanpur, The JHU multi-microphone multi-speaker asr system for the CHiME-6 challenge, arXiv preprint arXiv:2006.07898 (2020).
- [85] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, et al., The STC system for the CHiME-6 challenge, in: CHiME 2020 Workshop on Speech Processing in Everyday Environments, 2020.
- [86] T. Gao, J. Du, L.-R. Dai, C.-H. Lee, Densely connected progressive learning for lstm-based speech enhancement, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2018, pp. 5054–5058.
- [87] H. Erdogan, J. R. Hershey, S. Watanabe, J. Le Roux, Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2015, pp. 708–712.
- [88] P. C. Loizou, Speech enhancement: theory and practice, CRC press, 2013.
- [89] J. Heymann, L. Drude, R. Haeb-Umbach, Neural network based spectral mask estimation for acoustic beamforming, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2016, pp. 196–200.
- [90] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, J. Le Roux, Improved MVDR beamforming using single-channel mask prediction networks, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2016, pp. 1981–1985.
- [91] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, B.-H. Juang, Speech dereverberation based on variance-normalized delayed linear prediction, *IEEE Transactions on Audio, Speech, and Language Processing* 18 (2010) 1717–1731.
- [92] T. Yoshioka, T. Nakatani, Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening, *IEEE Transactions on Audio, Speech, and Language Processing* 20 (2012) 2707–2720.
- [93] L. Drude, J. Heymann, C. Boeddeker, R. Haeb-Umbach, NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing, in: Speech Communication; 13th ITG-Symposium, VDE, 2018, pp. 1–5.
- [94] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, F. Alleva, Recognizing over-

- lapped speech in meetings: A multichannel separation approach using neural networks, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 3038–3042.
- [95] C. Boeddecker, J. Heitkaemper, J. Schmalenstroer, L. Drude, J. Heymann, R. Haeb-Umbach, Front-end processing for the CHiME-5 dinner party scenario, in: Proceedings of CHiME 2018 Workshop on Speech Processing in Everyday Environments, 2018, pp. 35–40.
- [96] X. Xiao, N. Kanda, Z. Chen, T. Zhou, T. Yoshioka, Y. Zhao, G. Liu, J. Wu, J. Li, Y. Gong, Microsoft speaker diarization system for the voxceleb speaker recognition challenge 2020, arXiv preprint arXiv:2010.11458 (2020).
- [97] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, A. Zisserman, VoxSRC 2020: The second VoxCeleb speaker recognition challenge, arXiv preprint arXiv:2012.06867 (2020).
- [98] A. Itu, silence compression scheme for g. 729 optimized for terminals conforming to recommendation v. 70, ITU-T Recommendation G 729 (1996).
- [99] R. Chengalvarayan, Robust energy normalization using speech/nonspeech discriminator for german connected digit recognition, in: Sixth European Conference on Speech Communication and Technology, 1999.
- [100] K.-H. Woo, T.-Y. Yang, K.-J. Park, C. Lee, Robust voice activity detection algorithm for estimating noise spectrum, Electronics Letters 36 (2000) 180–181.
- [101] E. Nemer, R. Goubran, S. Mahmoud, Robust voice activity detection using higher-order statistics in the lpc residual domain, IEEE Transactions on Speech and Audio Processing 9 (2001) 217–231.
- [102] J. Sohn, N. S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE signal processing letters 6 (1999) 1–3.
- [103] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, P. Matějka, Developing a speech activity detection system for the darpa rats program, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2012, pp. 1969–1972.
- [104] T. Pfau, D. P. Ellis, A. Stolcke, Multispeaker speech activity detection for the icsi meeting recorder, in: IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU’01., IEEE, 2001, pp. 107–110.
- [105] R. Sarikaya, J. H. Hansen, Robust detection of speech activity in the presence of noise, in: Proceedings of the International Conference on Spoken Language Processing, volume 4, Citeseer, 1998, pp. 1455–8.
- [106] N. Ryant, M. Liberman, J. Yuan, Speech activity detection on youtube using deep neural networks., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2013, pp. 728–731.
- [107] S. Thomas, S. Ganapathy, G. Saon, H. Soltau, Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2014, pp. 2519–2523.
- [108] G. Gelly, J.-L. Gauvain, Optimization of rnn-based speech activity detection, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2017) 646–656.
- [109] J. Kim, M. Hahn, Voice activity detection using an adaptive context attention model, IEEE Signal Processing Letters 25 (2018) 1181–1185.
- [110] D. Haws, D. Dimitriadis, G. Saon, S. Thomas, M. Picheny, On the importance of event detection for asr, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2016.
- [111] S. Chen, P. Gopalakrishnan, et al., Speaker, environment and channel change detection and clustering via the bayesian information criterion, in: Proceedings DARPA broadcast news transcription and understanding workshop, volume 8, Virginia, USA, 1998, pp. 127–132.
- [112] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, Strategies for automatic segmentation of audio data, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 3, 2000, pp. 1423–1426.
- [113] M. A. Siegler, U. Jain, B. Raj, R. M. Stern, Automatic segmentation, classification and clustering of broadcast news audio, in: Proc. DARPA speech recognition workshop, volume 1997, 1997.
- [114] H. Gish, M.-H. Siu, J. R. Rohlicek, Segregation of speakers for speech recognition and speaker identification., in: icassp, volume 91, 1991, pp. 873–876.
- [115] J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, C. Wellekens, A speaker tracking system based on speaker turn detection for nist evaluation, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 2, IEEE, 2000, pp. 1177–1180.
- [116] R. Gangadharaiah, B. Narayanaswamy, N. Balakrishnan, A novel method for two-speaker segmentation, in: Eighth International Conference on Spoken Language Processing, 2004.
- [117] A. Tritschler, R. A. Gopinath, Improved speaker segmentation and segments clustering using the bayesian information criterion, in: Sixth European Conference on Speech Communication and Technology, 1999.
- [118] P. Delacourt, C. J. Wellekens, Distbic: A speaker-based segmentation for audio data indexing, Speech Communication 32 (2000) 111–126.
- [119] K. Mori, S. Nakagawa, Speaker change detection and speaker clustering using vq distortion for broadcast news speech recognition, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 1, 2001, pp. 413–416.
- [120] J. Ajmera, I. McCowan, H. Bourlard, Robust speaker change detection, IEEE signal processing letters 11 (2004) 649–651.
- [121] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, J. Fortuna, Unsupervised speaker change detection using probabilistic pattern matching, IEEE signal processing letters 13 (2006) 509–512.
- [122] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, S. Khudanpur, Diarization is hard: some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2018, pp. 2808–2812.
- [123] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted gaussian mixture models, Digital signal processing 10 (2000) 19–41.
- [124] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Speaker and session variability in gmm-based speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 15 (2007) 1448–1460.
- [125] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, A study of interspeaker variability in speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 16 (2008) 980–988.
- [126] P. Kenny, G. Boulianne, P. Dumouchel, Eigenvoice modeling with sparse training data, IEEE Transactions on Speech and Audio Processing 13 (2005) 345–354.
- [127] W. Zhu, J. Pelecanos, Online speaker diarization using adapted i-vector transforms, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2016, pp. 5045–5049.
- [128] A. Kanagasundaram, D. Dean, R. Vogt, M. McLaren, S. Sridharan, M. Mason, Weighted lda techniques for i-vector based speaker verification, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2012, pp. 4781–4784.
- [129] A. Kanagasundaram, D. Dean, S. Sridharan, M. McLaren, R. Vogt, i-vector based speaker recognition using advanced channel compensation techniques, Computer Speech & Language 28 (2014) 121–140.
- [130] M. Senoussaoui, P. Kenny, N. Dehak, P. Dumouchel, et al., An i-vector extractor suitable for speaker recognition with both microphone and telephone speech., in: Odyssey, 2010, p. 6.
- [131] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, M. Mason, i-vector based speaker recognition on short utterances, in: Proceedings of the 12th Annual Conference of the International Speech Communication Association, International Speech Communication Association, 2011, pp. 2341–2344.
- [132] P. Matějka, O. Glembek, F. Castaldo, M. J. Alam, O. Plhot, P. Kenny, L. Burget, J. Černocký, Full-covariance ubm and heavy-tailed plda in i-vector speaker verification, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2011, pp. 4828–4831.
- [133] D. Garcia-Romero, C. Y. Espy-Wilson, Analysis of i-vector Length Normalization in Speaker Recognition Systems, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2011, pp. 249–252.
- [134] P. Kenny, Bayesian speaker verification with heavy-tailed priors., in: Odyssey, volume 14, 2010.
- [135] Y. Jiang, K. A. Lee, L. Wang, Plda in the i-supervector space for text-independent speaker verification, EURASIP Journal on Audio, Speech,

- and Music Processing 2014 (2014) 1–13.
- [136] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.
  - [137] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.
  - [138] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, et al., State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 1488–1492.
  - [139] K. J. Han, S. S. Narayanan, A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2007.
  - [140] J. E. Rougui, M. Rza, D. Aboutajdine, M. Gelgon, J. Martinez, Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 5, IEEE, 2006.
  - [141] S. Novoselov, A. Gusev, A. Ivanov, T. Pekhovsky, A. Shulipa, A. Avdeeva, A. Gorlanov, A. Kozlov, Speaker diarization with deep speaker embeddings for dihard challenge ii., in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 1003–1007.
  - [142] U. Von Luxburg, A tutorial on spectral clustering, *Statist. and Comput.* 17 (2007) 395–416.
  - [143] T. J. Park, K. J. Han, M. Kumar, S. Narayanan, Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap, *IEEE Signal Processing Letters* 27 (2019) 381–385.
  - [144] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 14, Oakland, CA, USA, 1967, pp. 281–297.
  - [145] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, *Advances in neural information processing systems* 14 (2001) 849–856.
  - [146] H. Ning, M. Liu, H. Tang, T. S. Huang, A spectral clustering approach to speaker diarization, in: Proceedings of the International Conference on Spoken Language Processing, 2006, pp. 2178–2181.
  - [147] J. Luque, J. Hernando, On the use of agglomerative and spectral clustering in speaker diarization of meetings, in: *Odyssey*, 2012, pp. 130–137.
  - [148] Q. Lin, R. Yin, M. Li, H. Bredin, C. Barras, Lstm based similarity measurement with spectral clustering for speaker diarization, *Proc. Interspeech* 2019 (2019) 366–370.
  - [149] Z. Zajíc, M. Kunešová, V. Radová, Investigation of Segmentation in i-vector Based Speaker Diarization of Telephone Speech, in: *International Conference on Speech and Computer*, 2016, pp. 411–418.
  - [150] D. Dimitriadis, P. Fousek, Developing on-line speaker diarization system, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2017, pp. 2739–2743.
  - [151] D. Comaniciu, P. Meer, Mean shift: A robust approach toward feature space analysis, *IEEE Transactions on pattern analysis and machine intelligence* 24 (2002) 603–619.
  - [152] T. Stafylakis, V. Katsouros, G. Carayannis, Speaker clustering via the mean shift algorithm, *Recall* 2 (2010) 7.
  - [153] M. Senoussaoui, P. Kenny, P. Dumouchel, T. Stafylakis, Efficient iterative mean shift based cosine dissimilarity for multi-recording speaker clustering, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 7712–7715.
  - [154] I. Salmun, I. Shapiro, I. Opher, I. Lapidot, Plda-based mean shift speakers' short segments clustering, *Computer Speech and Language* 45 (2017) 411–436.
  - [155] P. Kenny, D. Reynolds, F. Castaldo, Diarization of telephone conversations using factor analysis, *IEEE Journal of Selected Topics in Signal Processing* 4 (2010) 1059–1070.
  - [156] M. Diez, L. Burget, F. Landini, J. Černocký, Analysis of speaker diarization based on Bayesian HMM with eigenvoice priors, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019) 355–368.
  - [157] J. G. Fiscus, A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE, 1997, pp. 347–354.
  - [158] N. Brummer, L. Burget, J. Cernocký, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, A. Strasheim, Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006, *IEEE Transactions on Audio, Speech, and Language Processing* 15 (2007) 2072–2084.
  - [159] M. Huijbregts, D. van Leeuwen, F. Jong, The majority wins: a method for combining speaker diarization systems, in: Proceedings of the Annual Conference of the International Speech Communication Association, ISCA, 2009, pp. 924–927.
  - [160] S. Bozonnet, N. Evans, X. Anguera, O. Vinyals, G. Friedland, C. Fiedouille, System output combination for improved speaker diarization, in: Proceedings of the Annual Conference of the International Speech Communication Association, ISCA, 2010, pp. 2642–2645.
  - [161] A. Stolcke, T. Yoshioka, DOVER: A method for combining diarization outputs, in: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE, 2019, pp. 757–763.
  - [162] F. Landini, J. Profant, M. Diez, L. Burget, Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks, *arXiv preprint arXiv:2006.07898* (2020).
  - [163] G. Sell, D. Garcia-Romero, Diarization resegmentation in the factor analysis subspace, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2015, pp. 4794–4798.
  - [164] M. Diez, L. Burget, F. Landini, S. Wang, H. Černocký, Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 6519–6523.
  - [165] Q. Lin, Y. Hou, M. Li, Self-attentive similarity measurement strategies in speaker diarization, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2020, pp. 284–288.
  - [166] T. J. Park, M. Kumar, S. Narayanan, Multi-scale speaker diarization with neural affinity score fusion, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, pp. 7173–7177.
  - [167] N. Ftomotomos, D. Dimitriadis, A memory augmented architecture for continuous speaker identification in meetings, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 6524–6528.
  - [168] S. Horiguchi, P. Garcia, Y. Fujita, S. Watanabe, K. Nagamatsu, End-to-end speaker diarization as post-processing, *arXiv preprint arXiv:2012.10055* (2020).
  - [169] K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, Tackling real noisy reverberant meetings with all-neural source separation, counting, and diarization system, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 381–385.
  - [170] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, K. Nagamatsu, End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2020, pp. 269–273.
  - [171] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, J. Shi, K. Nagamatsu, Neural speaker diarization with speaker-wise chain rule, *arXiv preprint arXiv:2006.01796* (2020).
  - [172] Y. Xue, S. Horiguchi, Y. Fujita, S. Watanabe, K. Nagamatsu, Online end-to-end neural diarization with speaker-tracing buffer, *arXiv preprint arXiv:2006.02616* (2020).
  - [173] E. Han, C. Lee, A. Stolcke, BW-EDA-EEND: Streaming end-to-end neural speaker diarization for a variable number of speakers, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, pp. 7193–7197.
  - [174] K. Kinoshita, M. Delcroix, N. Tawara, Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, pp. 7198–7202.
  - [175] N. Kanda, Y. Gaur, X. Wang, Z. Meng, Z. Chen, T. Zhou, T. Yoshioka,

- Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2020, pp. 36–40.
- [176] N. Kanda, Z. Meng, L. Lu, Y. Gaur, X. Wang, Z. Chen, T. Yoshioka, Minimum Bayes risk training for end-to-end speaker-attributed ASR, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 6503–6507.
- [177] E. Ustinova, V. Lempitsky, Learning deep embeddings with histogram loss, *Proceedings of Advances in Neural Information Processing Systems* 29 (2016) 4170–4178.
- [178] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: *Proceedings of International Conference on Machine Learning*, 2016, pp. 478–487.
- [179] X. Guo, L. Gao, X. Liu, J. Yin, Improved deep embedded clustering with local structure preservation, in: *Proceedings of International Joint Conference on Artificial Intelligence*, 2017, pp. 1753–1759.
- [180] Y. LeCun, Y. Bengio, G. Hinton, *Deep Learning*, *Nature* 521 (2015) 436.
- [181] A. Santoro, R. Faulkner, D. Raposo, J. Rae, M. Chrzanowski, T. Weber, D. Wierstra, O. Vinyals, R. Pascanu, T. Lillicrap, Relational Recurrent Neural Networks, in: *Proceedings of Advances in Neural Information Processing Systems*, 2018, pp. 7299–7310.
- [182] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, Meta-learning with Memory-Augmented Neural Networks, in: *Proceedings of International Conference on Machine Learning*, 2016, pp. 1842–1850.
- [183] S. Sukhbaatar, J. Weston, R. Fergus, et al., End-to-End Memory Networks, in: *Proceedings of Advances in Neural Information Processing Systems*, 2015, pp. 2440–2448.
- [184] T. Yoshioka, D. Dimitriadis, A. Stolcke, W. Hinthorn, Z. Chen, M. Zeng, H. Xuedong, Meeting Transcription Using Asynchronous Distant Microphones, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2019, pp. 2968–2972.
- [185] D. M. Blei, P. I. Frazier, Distance dependent chinese restaurant processes., *Journal of Machine Learning Research* 12 (2011).
- [186] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2016) 1137–1149.
- [187] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, R. Horaud, An EM algorithm for joint source separation and diarisation of multichannel convolutive speech mixtures, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2017, pp. 16–20.
- [188] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, R. Horaud, S. Gannot, Exploiting the intermittency of speech for joint separation and diarization, in: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, 2017, pp. 41–45.
- [189] K. Maekawa, Corpus of spontaneous japanese: Its design and evaluation, in: *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 7–12.
- [190] J. Huang, E. Marcheret, K. Visweswariah, G. Potamianos, The ibm rt07 evaluation systems for speaker diarization on lecture meetings, in: *Multimodal Technologies for Perception of Humans*, Springer, 2007, pp. 497–508.
- [191] L. Canseco-Rodriguez, L. Lamel, J.-L. Gauvain, Speaker diarization from speech transcripts, in: *Proceedings of the International Conference on Spoken Language Processing*, volume 4, 2004, pp. 3–7.
- [192] N. Flemotomos, P. Georgiou, S. Narayanan, Linguistically aided speaker diarization using speaker role information, in: *Odyssey*, 2020, pp. 117–124.
- [193] T. J. Park, K. J. Han, J. Huang, X. He, B. Zhou, P. Georgiou, S. Narayanan, Speaker diarization with lexical information, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2019, pp. 391–395.
- [194] J. Fiscus, J. Ajot, J. Garofolo, The Rich Transcription 2007 meeting recognition evaluation, 2007, pp. 373–389.
- [195] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, T. Nakatani, Speaker-aware neural network based beamformer for speaker extraction in speech mixtures., in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2017, pp. 2655–2659.
- [196] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, T. Nakatani, Single channel target speaker extraction and recognition with speaker beam, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2018, pp. 5554–5558.
- [197] M. Delcroix, S. Watanabe, T. Ochiai, K. Kinoshita, S. Karita, A. Ogawa, T. Nakatani, End-to-end SpeakerBeam for single channel target speech recognition., in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2019, pp. 451–455.
- [198] N. Kanda, Y. Gaur, X. Wang, Z. Meng, T. Yoshioka, Serialized output training for end-to-end overlapped speech recognition, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2020, pp. 2797–2801.
- [199] N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng, Z. Chen, T. Yoshioka, End-to-end speaker-attributed asr with transformer, *arXiv preprint arXiv:2104.02128* (2021).
- [200] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al., The ami meeting corpus: A pre-announcement, in: *International workshop on machine learning for multimodal interaction*, Springer, 2005, pp. 28–39.
- [201] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Pelskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, The ICSI meeting corpus, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003, pp. 1–364–1–367.
- [202] J. Barker, S. Watanabe, E. Vincent, J. Trmal, The fifth ‘chime’ speech separation and recognition challenge: Dataset, task and baselines, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2018, pp. 1561–1565.
- [203] J. S. Chung, A. Nagrani, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, A. Zisserman, VoxSRC 2019: The first VoxCeleb speaker recognition challenge, *arXiv preprint arXiv:1912.02522* (2019).
- [204] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, A. Zisserman, Spot the conversation: Speaker diarisation in the wild, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2020, pp. 299–303.
- [205] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, LibriSpeech: an ASR corpus based on public domain audio books, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2015, pp. 5206–5210.
- [206] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The first dihard speech diarization challenge, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2018.
- [207] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, et al., Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario, *arXiv preprint arXiv:2104.03603* (2021).
- [208] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, K. Choukri, The ester evaluation campaign for the rich transcription of french broadcast news., in: *LREC*, 2004.
- [209] S. Galliano, G. Gravier, L. Chaubard, The ester 2 evaluation campaign for the rich transcription of french radio broadcasts, in: *Tenth Annual Conference of the International Speech Communication Association*, 2009.
- [210] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, O. Galibert, The etape corpus for the evaluation of speech-based tv content processing in the french language, in: *LREC-Eighth international conference on Language Resources and Evaluation*, 2012, p. na.
- [211] C.-C. Chiu, A. Tripathi, K. Chou, C. Co, N. Jaitly, D. Jaunzeikare, A. Kannan, P. Nguyen, H. Sak, A. Sankar, et al., Speech recognition for medical conversations, *arXiv preprint arXiv:1711.07274* (2017).
- [212] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. P. andD. Reidsma, P. Wellner, The AMI meeting corpus: a pre-announcement, in: *Proceedings of Int. Worksh. Machine Learning for Multimodal Interaction*, 2006, pp. 28–39.
- [213] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig, Achieving human parity in conversational speech recognition, *arXiv preprint arXiv:1610.05256* (2016).
- [214] G. Saon, G. Kurata, T. Seru, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, et al., English conver-



- sational telephone speech recognition by humans and machines, arXiv preprint arXiv:1703.02136 (2017).
- [215] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. Fabian, M. Espi, T. Higuchi, S. Araki, T. Nakatani, The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices, in: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, pp. 436–443.
  - [216] J. Du, Y. Tu, L. Sun, F. Ma, H. Wang, J. Pan, C. Liu, J. Chen, C. Lee, The USTC-iFlytek system for CHiME-4 challenge, in: *Proceedings of CHiME-4 Workshop*, 2016, pp. 36–38.
  - [217] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafan, H. Sak, G. Punduk, K. Chin, K. C. Sim, R. J. Weiss, K. W. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintrauba, E. McDermott, R. Rose, M. Shannon, Acoustic modeling for Google Home, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2017, pp. 399–403.
  - [218] D. Dimitriadis, P. Fousek, Developing on-line speaker diarization system, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2017, pp. 2739–2743.
  - [219] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
  - [220] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
  - [221] T. Yoshioka, I. Abramovski, C. Aksoylar, Z. Chen, M. David, D. Dimitriadis, Y. Gong, I. Gurvich, X. Huang, Y. Huang, A. Hurvitz, L. Jiang, S. Koubi, E. Krupka, I. Leichter, C. Liu, P. Parthasarathy, A. Vinnikov, L. Wu, X. Xiao, W. Xiong, H. Wang, Z. Wang, J. Zhang, Y. Zhao, T. Zhou, Advances in Online Audio-Visual Meeting Transcription, in: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2019, pp. 276–283.
  - [222] H. Buchner, R. Aichner, W. Kellermann, A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics, *IEEE Transactions on Speech and Audio Processing* 13 (2005) 120–134.
  - [223] H. Sawada, S. Araki, S. Makino, Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS, in: *Int. Symp. Circ., Syst., 2007*, pp. 3247–3250.
  - [224] F. Nesta, P. Svaizer, M. Omologo, Convolutional bss of short mixtures by ica recursively regularized across frequencies, *IEEE Transactions on Audio, Speech, and Language Processing* 19 (2011) 624–639.
  - [225] H. Sawada, S. Araki, S. Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment, *IEEE Transactions on Audio, Speech, and Language Processing* 19 (2011) 516–527.
  - [226] N. Ito, S. Araki, T. Yoshioka, T. Nakatani, Relaxed disjointness based clustering for joint blind source separation and dereverberation, in: *Proceedings of International Workshop on Acoustic Echo and Noise Control*, 2014, pp. 268–272.
  - [227] L. Drude, R. Haeb-Umbach, Tight integration of spatial and spectral features for BSS with deep clustering embeddings, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2017, pp. 2650–2654.
  - [228] M. Maciejewski, G. Sell, L. P. Garcia-Perera, S. Watanabe, S. Khudanpur, Building corpora for single-channel speech separation across multiple domains, arXiv preprint arXiv:1811.02641 (2018).
  - [229] S. Araki, N. Ono, K. Kinoshita, M. Delcroix, Meeting recognition with asynchronous distributed microphone array using block-wise refinement of mask-based MVDR beamformer, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5694–5698.
  - [230] A. Stolcke, Making the most from multiple microphones in meeting recordings, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4992–4995.
  - [231] S. Narayanan, P. G. Georgiou, Behavioral signal processing: Deriving human behavioral informatics from speech and language, *Proceedings of the IEEE* 101 (2013) 1203–1233.
  - [232] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, S. Narayanan, Signal processing and machine learning for mental health research and clinical applications, *IEEE Signal Processing Magazine* 34 (2017) 189–196.
  - [233] M. Kumar, S. H. Kim, C. Lord, S. Narayanan, Speaker diarization for naturalistic child-adult conversational interactions using contextual information., *Journal of the Acoustical Society of America* 147 (2020) EL196–EL200. doi:10.1121/10.0000736.
  - [234] P. G. Georgiou, M. P. Black, S. S. Narayanan, Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments, in: *Proceedings of the joint ACM workshop on Human gesture and behavior understanding*, 2011, pp. 7–12.
  - [235] B. Xiao, C. Huang, Z. E. Imel, D. C. Atkins, P. Georgiou, S. S. Narayanan, A technology prototype system for rating therapist empathy from audio recordings in addiction counseling, *PeerJ Computer Science* 2 (2016) e59.
  - [236] S. N. Chakravarthula, M. Nasir, S.-Y. Tseng, H. Li, T. J. Park, B. Baucum, C. J. Bryan, S. Narayanan, P. Georgiou, Automatic prediction of suicidal risk in military couples using multimodal interaction cues from couples conversations, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6539–6543.
  - [237] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, H. Christensen, Toward the automation of diagnostic conversational analysis in patients with memory complaints, *Journal of Alzheimer’s Disease* 58 (2017) 373–387.
  - [238] G. P. Finley, E. Edwards, A. Robinson, N. Sadoughi, J. Fone, M. Miller, D. Suendermann-Oeft, M. Brenndorfer, N. Axtmann, An automated assistant for medical scribes., in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2018, pp. 3212–3213.
  - [239] A. Guo, A. Faria, J. Riedhammer, Remeeting – Deep insights to conversations, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2016, pp. 1964–1965.
  - [240] A. Adhese, Y. Yu, A. Eshghi, A comprehensive evaluation of incremental speech recognition and diarization for conversational ai, in: *Proceedings of the International Conference on Computational Linguistics*, 2020, pp. 3492–3503.
  - [241] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, M. Liberman, Third dihard challenge evaluation plan, arXiv preprint arXiv:2006.05815 (2020).
  - [242] O. Cetin, E. Shriberg, Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, IEEE, 2006, pp. 357–360.
  - [243] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, R. Haeb-Umbach, Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR, in: *Proceedings of the Annual Conference of the International Speech Communication Association*, 2019, pp. 1248–1252.
  - [244] J. S. Garofolo, C. D. Laprun, J. G. Fiscus, The rich transcription 2004 spring meeting recognition evaluation, NIST, 2004.
  - [245] S. Otterson, M. Ostendorf, Efficient use of overlap information in speaker diarization, in: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, IEEE, 2007, pp. 683–686.
  - [246] K. Boakye, B. Trueba-Hornero, O. Vinyals, G. Friedland, Overlapped speech detection for improved speaker diarization in multiparty meetings, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2008, pp. 4353–4356.
  - [247] L. Bullock, H. Bredin, L. P. Garcia-Perera, Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection, in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2020, pp. 7114–7118.